Information Theoretic Regression Methods

Dedicated to the memory of Solomon Kullback, 1907-1994

Ehsan Soofi

Technical Report No. 126
April, 1996

Center for Computational Statistics

> *"We shall use information in the technical sense to be defined, and it should not be confused with our semantic concept, though it is true that the properties of the measure of information following from the technical definition are such as to be reasonable according to our intuitive notion of information."*
> Kullback (1959)

George Mason University
Fairfax, VA 22030

CENTER FOR COMPUTATIONAL STATISTICS
TECHNICAL REPORT SERIES

TTR 110. Edward J. Wegman, Huge Data Sets and the Frontiers of Computational Feasibility, November, 1994. published ***Journal of Computational and Graphical Statistics***, 4(4), 281-195, 1995.

TR 111. Winston C. Chow, Fractional Process Modeling, November, 1994.

TR 112. Mark C. Sullivan, ***Computationally Efficient Statistical Signal Processing Using Nonlinear Operators*** (Ph.D. Dissertation), December, 1994.

TR 113. Irwin Greenberg, Some Simple Approximation Methods in Level Crossing Problems, December, 1994.

TR 114. Jeffrey L. Solka, Wendy L. Poston and Edward J. Wegman, A New Visualization Technique to Study the Time Evolution of Finite and Adaptive Mixture Estimators, December, 1994. published ***Journal of Computational and Graphical Statistics***, 4(3), 180-198, 1995.

TR 115. D. B. Carr and A. R. Olsen, Representing Cumulative Distributions with Parallel Coordinate Plots, August, 1995

TR 116. Jeffrey L. Solka, ***Matching Model Information Content to Data Information*** (Ph.D. Dissertation), August, 1995.

TR 117. Wendy L. Poston, ***Optimal Subset Selection Methods***, (Ph.D. Dissertation), August, 1995.

TR 118. Clifton D. Sutton, Sphere Packing, August, 1995.

TR 119. Wendy L. Poston, Edward J. Wegman, and Jeffrey L. Solka, A Parallel Algorithm for Subset Selection, August, 1995.

TR 120. Barnabas Takacs, Harry Wechsler, and Edward J. Wegman, A Model of Active Perception and its Implementation on the Intel Paragon XP/S, August, 1995.

TR 121. Shan-chuan Li, Walter Dyar, and Mary-Ellen Verona, GRASS Database Explored and Applied to Biodiversity Query with Splus, August, 1995, to appear ***Computing Science and Statistics***, 27, 1995.

TR 122. Kathleen Golitko Perez-Lopez, ***Management of Scientific Image Databases Using Wavelets*** (Ph.D. Dissertation), August, 1995.

TR 123. Edward J. Wegman, Jeffrey L. Solka and Wendy L. Poston, Immersive Methods for Mine Warfare, April, 1996.

TR 124. Edward J. Wegman and Qiang Luo, High Dimensional Clustering using Parallel Coordinates and the Grand Tour, April, 1996.

TR 125. Kletus A. Lawler, ***Linear and Nonlinear Regression Estimates for a Cobb-Douglas Model***, (M.S. Thesis), April, 1996.

TR 126. Ehsan S. Soofi, Information Theoretic Regression Methods, April, 1996.

INFORMATION THEORETIC REGRESSION METHODS

Dedicated to the memory of Solomon Kullback 1907-1994

Ehsan S. Soofi

School of Business Administration
University of Wisconsin-Milwaukee
P.O. Box 742, Milwaukee, WI 53201

Completed in Spring 1996 while the author was visiting
Center for Computational Statistics
George Mason University
Fairfax, VA 22030-4444

To appear in

Advances in Econometrics: Applying Maximum Entropy to Econometric Problems,

Vol. 12, 1996, eds., T. B. Fomby and R. C. Hill, JAI Press.

DATE OF DRAFT: April 1996

Contents

1 Introduction

Since the publication of the seminal note, Kullback and Leibler (1951), there has been continual endeavor in statistics and related fields to explicate the existing statistical methods and to develop new methods based on the *logarithmic information* of Shannon (1948). There are many fine collections of information-theoretic methodologies and their applications to the related fields such as Kullback (1954, 1959), Lindley (1956), Jaynes (1957, 1968, 1982), Theil (1967), Akaike (1973), Gokhale and Kullback (1978), Shore and Johnson (1980), Kapur (1989), Brockett (1991), Cover and Thomas (1991), Csiszar (1991), Zellner (1991), Maasoumi (1993), and Soofi (1994).

During the last four decades numerous information theoretic regression methods have been developed. Kullback and Rosenblatt (1957) pioneered the information theoretic approach to regression by explicating the usual regression quantities such as sums of squares and F-ratios in terms of information functions. We have now information theoretic methods for model and predictive density derivation, parameter estimation and testing, model selection, collinearity analysis, and influential observation detection which can be used in sampling theory and Bayesian regression analyses.

The logical foundation, elegance, and versatility of the information theoretic approach have been increasingly attracting the attention of researchers in various fields. However, the available entropy-based methods are not yet commonly used in the mainstream regression analysis. Many information-theoretic regression methods are developed disjointly in the context of providing alternatives to particular problems rather than as integral parts of a system of regression analysis. Information-theoretic interpretation of many of the available methods and the relationship among them have not yet been fully explicated. The purpose of this paper is to integrate the existing entropy-based methods in a single framework, to explore their interrelationships, to elaborate on information theoretic interpretations of the existing entropy-based diagnostics, and to present information theoretic interpretations for some traditional diagnostics.

2 Information Functions

In this section, the basic information functions used in regression analysis, their properties, interpretations, and relationships with the Fisher information are reviewed.

2.1 Entropy

The entropy of a continuous random variable X is defined as

$$H(X) \equiv H[f(x)] = -\int_{-\infty}^{\infty} f(x)\ log\ f(x)dx,$$

where $f(x)$ is the probability density function for the absolutely continuous distribution F.

The differential entropy may be negative or infinite. Boundedness of $f(x)$ implies $H(X) > -\infty$ (Ash 1965, p. 237). For a distribution with finite variance, the entropy is finite, but the converse may not hold.

The conditional entropy is obtained by using the conditional density in the entropy expression, $H(Y|x) = H[f(y|x)]$. A conditioning may increase or decrease the entropy. The expected conditional entropy is defined by

$$H(Y|X) \equiv E_x[H(Y|x)] \leq H(Y);$$

the equality holds if and only if the two variables are independent. That is, on average, conditioning decreases the entropy.

The entropy of an n-dimensional random variable $\boldsymbol{X} = (X_1, \cdots, X_n)'$ is obtained by using the joint density $f(\boldsymbol{x})$ in the entropy expression. For the joint entropy of an n-dimensional random variable, we have

$$H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1) \leq \sum_{i=1}^{n} H(X_i).$$

In the last relation, the equality holds if and only if the random variables $X_1, \cdots, X_n$ are independent.

The differential entropy is not invariant under one-to-one transformations of X. For any continuous random variable $Z = g(X)$,

$$H(Z) = H(X) - E\left[log\left|\frac{d}{dZ}g^{-1}(Z)\right|\right]. \tag{2.1}$$

A random variable X with a location parameter μ and scale parameter σ may be written as $X = \sigma Z + \mu$, where the distribution of Z is independent of μ and σ. Using (2.1), we find

$$H(X|\mu,\sigma) = H(Z) + log\sigma.$$

Thus the entropy is location invariant but not scale invariant.

$H(f)$ is concave in f. The Maximum Entropy (ME) model $f^*(x|\boldsymbol{\theta})$ is the density that maximizes $H[f(x|\theta)]$ subject to the information moment constraints,

$$E_f[c_m(X)|\boldsymbol{\theta}] = \theta_m, \; m = 1, \cdots, M, \tag{2.2}$$

where c_m's are integrable with respect to f and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_M)$ is the vector of moment values. The moment values might be known quantities (e.g., computed from the data) or unknown parameters. The ME solution, if it exists, is in the form of

$$f^*(x|\boldsymbol{\theta}) = C(\boldsymbol{\theta})e^{\eta_1 c_1(x) + \cdots + \eta_M c_M(x)}, \tag{2.3}$$

where $C(\boldsymbol{\theta})$ is the normalizing constant for the ME density and $\eta_m = \eta_m(\boldsymbol{\theta}), m = 1, \cdots, M$ are Lagrange multipliers for enforcing the information constraints (2.2). Multivariate ME distributions are found similarly.

The entropy measures the "uniformity" of a distribution and provides a measure of information in the following sense. $H(X)$ increases as the concentration of probabilities over subsets of the support of the distribution decreases. This feature makes $H(X)$ a suitable measure of *uncertainty* associated with $f(x)$. The term *uncertainty* describes the difficulty of predicting an outcome x of a random variable X with the probability distribution $f(x)$. A distribution $f_1(x)$ with a large entropy is less concentrated (more difficult to predict its outcomes) than a distribution $f_2(x)$ with a smaller entropy. Thus, $f_1(x)$ is less informative as compared with $f_2(x)$. Some authors have interpreted $-H(X)$ as an information criterion in the context of developing least informative probability distribution; see, e.g., Zellner 1971)

The ME distribution is the least informative distribution since it does not include any information that is not explicitly formulated as a constraint in (2.2). The information content of each moment constraint in (2.2) is reflected in the uncertainty reduction power of that constraint (Jaynes 1968, 1982; Soofi 1992, 1994). Suppose that the ME distributions exit for all M constraints in (2.2) and for a subset of ℓ constraints, $\ell < M$. Then the amount of information provided by the additional constraints $E_f[c_{\ell+1}(X)|\boldsymbol{\theta}] = \theta_{\ell+1}, \cdots, E_f[c_M(X)|\boldsymbol{\theta}] = \theta_M$, is quantified by the amount of entropy reduction,

$$\Delta H(f^*_{1,\cdots,\ell}, f^*_{1,\cdots,M}) = H(f^*_{1,\cdots,\ell}) - H(f^*_{1,\cdots,M}) \geq 0, \quad 0 \leq \ell \leq M.$$

Information indices are constructed by mapping the entropy reduction to the unit interval.

The *information index* of $M - \ell$ additional constraints on a continuous ME distribution may be computed by the following exponential transformation of the entropy reduction

$$I^*_C(c_{\ell+1}, \cdots, c_M) = 1 - e^{-\Delta H(f^*_{1,\ldots,\ell}, f^*_{1,\ldots,M})}, \quad 0 \leq \ell \leq M.$$

An $I^*_C(c_{\ell+1}, \cdots, c_M) \approx 0$ indicates that the additional constraints are redundant for concentrating the probabilities. An $I^*_C(c_{\ell+1}, \cdots, c_M) \approx 1$ indicates that the first set of constraints are redundant. In particular, for $\ell = 0$, the ME distribution over an infinite support is improper uniform with infinite entropy, thus, $I^*_C(c_1, \cdots, c_M) = 1$.

An information index of distributions in a specific class is defined in Section 2.2.

Example 2.1

(i) The n-variate ME density subject to the information constraints

$$E(\boldsymbol{X}) = \boldsymbol{\mu}, \qquad E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = \Sigma \tag{2.4}$$

is the n-variate normal $N(\boldsymbol{\mu}, \Sigma)$. The entropy of $N(\boldsymbol{\mu}, \Sigma)$ is

$$H(\boldsymbol{X}) = \frac{n}{2} log(2\pi e) + \frac{1}{2} log\ |\Sigma|,$$

where $|\Sigma|$ denotes the determinant of the covariance matrix.

(ii) Consider the following constraints for a bivariate random variable, $\boldsymbol{X}$:

$$c_i(\boldsymbol{X}) = X_i^2, \qquad E(X_i^2) = \sigma^2, \quad i = 1, 2.$$

The ME distribution subject to these constraints is the bivariate normal $f^*(\boldsymbol{x}; \sigma^2) = N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_2)$, where $\boldsymbol{I}_n$ is the identity matrix of order n. Note that since the constraint $c_i(\boldsymbol{X})$, $i = 1, 2$ only include information about the marginal moments, the ME solution is independence between the components of $\boldsymbol{X}$. That is, when the information about a relationship between the components of a random variable is not present in the information constraints, the ME solution reflects the absence of a relationship.

(iii) Now consider the additional cross-product constraint

$$c_3(\boldsymbol{X}) = X_1X_2, \qquad E(X_1X_2) = \rho\sigma^2.$$

The ME distribution subject to c_1, c_2 and c_3 is the bivariate normal distribution

$$f^*(\boldsymbol{x};\sigma^2,\ \rho) = N\left(\mathbf{0}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

Hence, the ME solution reflects the information specified in terms of the correlation in the cross-product constraint.

The entropy reduction due to c_3 is

$$H[f^*(\boldsymbol{X};\sigma^2)] - H[f^*(\boldsymbol{X};\sigma^2,\ \rho)] = -\frac{1}{2}log(1-\rho^2) \geq 0.$$

The partial information index of the additional constraint is

$$I^*_C(X_1X_2) = 1-(1-\rho^2)^{1/2}.$$

Thus, for example, for $\rho = .6$ the uncertainty reduction is 20% and for $\rho = .8$ the uncertainty reduction is 40%.

2.2 Discrimination Information

The most widely known information theoretic measure of discrepancy between two distributions is the Kullback-Leibler discrimination information function

$$\begin{aligned} K(f:g) &= \int_{-\infty}^{\infty} f(x)\ log\ \frac{f(x)}{g(x)}dx \\ &= -H[f(x)] - E_f\{log[g(X)]\}. \end{aligned} \tag{2.5}$$

The discrimination information function between two multivariate distributions is defined similarly. $K(f:g)$ is well-defined as long as $g(x)=0$ only if $f(x)=0$.

$K(f : g)$ is the entropy of F relative to G. It is also referred to as the cross-entropy between the two distributions. In general, there is no relationship between $K(f : g)$ and $H(g)$.

If $f(x|\boldsymbol{\theta})$ is a distribution in the class $\Omega_{\boldsymbol{\theta}}$ of distributions that satisfy (2.2) and $f^*(x|\boldsymbol{\theta})$ is the ME distribution in $\Omega_{\boldsymbol{\theta}}$, then (Soofi, Ebrahimi, and Habibullah 1995)

$$K(f : f^*|\boldsymbol{\theta}) = H[f^*(x|\boldsymbol{\theta})] - H[f(x|\boldsymbol{\theta})].$$

The *Information Discrimination (ID)* index of a distribution in $\Omega_{\boldsymbol{\theta}}$ is defined by

$$ID(f : f^*|\boldsymbol{\theta}) = 1 - e^{-K(f:f^*|\boldsymbol{\theta})}.$$

A distribution $f \in \Omega_{\boldsymbol{\theta}}$ is said to be *ID distinguishable* with the ME model if

$$ID(f : f^*|\boldsymbol{\theta}) > ID(f^* : f^*|\boldsymbol{\theta}) = 0. \tag{2.6}$$

The properties of $K(f : g)$ for discrete and continuous distributions are the same. Some properties of $K(f : g)$ are as follows (Kullback 1959):

(i) $K(f : g) \geq 0$; the equality holds if and only if $f(x) = g(x)$ almost everywhere.

(ii) For mutually independent random variables $X_1, \cdots, X_n$,

$$K[f(x_1, \cdots, x_n) : g(x_1, \cdots, x_n)] = \sum_{i=1}^{n} K[f(x_i) : g(x_i)].$$

(iii) For any two random variables X and Y,

$$\begin{aligned} K[f(x,y) : g(x,y)] &= K[f(x) : g(x)] + E_x\{K[f(y|x) : g(y|x)]\} \\ &= K[f(y) : g(y)] + E_y\{K[f(x|y) : g(x|y)]\}. \end{aligned}$$

Thus, for example, $K[f(x,y) : g(x,y)] \geq K[f(x) : g(x)]$; the equality holds if and only if the expected discrimination information between the respective conditional distributions is zero.

(iv) Let $Y = T(X)$ be a transformation and let $f_Y(y)$ and $g_Y(y)$ denote the distributions induced by T on $f_X(x)$ and $g_X(x)$. Then $K(f_Y : g_Y) \leq K(f_X : g_X)$ with equality if and only if

$$\frac{f_Y(T(x))}{g_Y(T(x))} = \frac{f_X(x)}{g_X(x)}, \tag{2.7}$$

for almost all x. If condition (2.7) holds, T is a *sufficient statistic for discrimination.*

(v) When $f(x;\theta)$ and $g(x;\theta)$, $K(f_Y : g_Y) \leq K(f : g)$ with equality if and only if T is a sufficient statistic for θ.

(vi) $K(f : g)$ is convex in f and in g.

The Minimum Discrimination Information (MDI) model reference to a distribution g is obtained by minimizing $K(f : g)$ with respect to f subject to the information constraint (2.2). The MDI density, if it exists, is given by

$$f^*(x; g, \boldsymbol{\theta}) = C(\boldsymbol{\theta})g(x)e^{\eta_1 c_1(x)+\cdots+\eta_M c_M(x)}.$$

When $K(f : g) = K(f : g; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the unknown parameter of one of the distributions. The parameter may be estimated by an MDI procedure; see Section 4.

<u>**Example 2.2**</u>

(i) The discrimination information between two n-dimensional normal distributions $f = N(\boldsymbol{\mu}_f, \Sigma_f)$ and $g = N(\boldsymbol{\mu}_g, \Sigma_g)$ is

$$K(f : g) = \frac{1}{2}[Tr(\Sigma_f\Sigma_g^{-1} - log|\Sigma_f\Sigma_g^{-1}| - n] + \frac{1}{2}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)'\Sigma_g^{-1}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g), \qquad (2.8)$$

where Tr denotes the trace of a matrix. The first term in (2.8) gives the information discrepancy due to two different covariance structures and the second term gives the information discrepancy due to two different means. For $\Sigma_f = \sigma_f^2\boldsymbol{I}_n$ and $\Sigma_g = \sigma_g^2\boldsymbol{I}_n$, (2.8) gives

$$K(f : g) = \frac{n}{2}\left[\frac{\sigma_f^2}{\sigma_g^2} - log\left(\frac{\sigma_f^2}{\sigma_g^2}\right) - 1\right] + \frac{(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)'(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)}{2\sigma_g^2}. \qquad (2.9)$$

(ii) Let $g(\boldsymbol{y}; \boldsymbol{\mu}_g, \Sigma_g) = N(\boldsymbol{\mu}_g, \Sigma_g)$ be an n-variate normal density. Then the MDI distribution reference to g subject to the mean information constraint

$$E_f(\boldsymbol{Y}) = \boldsymbol{\mu}_f \qquad (2.10)$$

is the n-variate normal $f^* = N(\boldsymbol{\mu}_f, \Sigma_g)$; proof is given in Soofi (1985). Thus, the minimum information discrepancy between the class of distributions that satisfy (2.10) and $N(\boldsymbol{\mu}_g, \Sigma_g)$is given by the second term in (2.8).

2.3 Mutual Information

The entropy difference, $\vartheta(Y|x) = H(Y) - H(Y|x)$ measures the information provided by the value x about the random variable Y. A particular value of x may or may not be informative which is indicated by the sign of $\vartheta(Y|x)$. The *mutual information* between two random variables is defined by

$$\begin{aligned} \vartheta(Y \wedge X) &\equiv E_x[\vartheta(Y|x)] \\ &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned} \tag{2.11}$$

In terms of the discrimination information, the mutual information is given by

$$\begin{aligned} \vartheta(Y \wedge X) &= K[f(x,y) : f(x)f(y)] \qquad (2.12) \\ &= E_x\{K[f(y|x) : f(y)]\} \\ &= E_y\{K[f(x|y) : f(x)]\}. \end{aligned}$$

Thus $\vartheta(Y \wedge X) = \vartheta(X \wedge Y) \geq 0$ with equality if and only if $f(x,y) = f(x)f(y)$. Accordingly, $\vartheta(Y \wedge X)$ is a measure of stochastic dependency between the two variables.

A useful normalization of $\vartheta(Y \wedge X)$ for the continuous case is

$$I_C(Y \wedge X) = 1 - e^{-2\vartheta(Y \wedge X)}.$$

$I_C(Y \wedge X)$ is an index of functional relationship between the two variables. It generalizes the correlation coefficient; see Example 2.3, part (iii). An $I_C(Y \wedge X) = 0$ indicates that two variables are independent. An $I_C(Y \wedge X) = 1$ indicates that the two variables are functionally dependent; see Joe (1989) for details.

$\vartheta(Y \wedge X)$ is invariant under one-to-one transformations of each variable. But $\vartheta(Y \wedge X)$ is not invariant under rotation of the coordinate system because (2.7) does not generally hold under rotations; see Example 2.3, part (ii).

In multivariate case, various mutual information functions may be obtained. The mutual information between the components of a p-dimensional random variable $\boldsymbol{X} = (X_1, \cdots, X_p)'$

is found by the multivariate extensions of (2.12) or (2.11) as:

$$\begin{aligned}\vartheta(\boldsymbol{X} \wedge X_1, \cdots, X_p) &= K[f(\boldsymbol{x}) : f(x_1) \cdots f(x_p)] \\ &= \sum_{j=1}^{p} H(X_j) - H(X_1, \cdots X_p).\end{aligned}$$

Mutual information functions for measuring other types of multivariate dependencies are found similarly.

The mutual information between a random variable Y and a p-dimensional random vector $\boldsymbol{X}$ is given by

$$\begin{aligned}\vartheta(Y \wedge \boldsymbol{X}) &= K[f(y, \boldsymbol{x}) : f(y) f(\boldsymbol{x})] \\ &= \sum_{j=1}^{p} \vartheta(Y \wedge X_j | X_{j-1}, \cdots, X_1).\end{aligned} \tag{2.13}$$

The *partial mutual information* function $\vartheta(Y \wedge X_j | X_{j-1}, \cdots, X_1)$ measures the conditional dependency between the pair (Y, X_j) given $X_1, \cdots, X_{j-1}$. In general, the decomposition (2.13) depends on the order of the variables $1, \cdots, p$. The partial mutual information may be interpreted as a measure of *relative importance* of X_j in a given order.

Example 2.3

(i) If $\boldsymbol{X} = (X_1, \cdots, X_p)'$ has multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$, then

$$\begin{aligned}\vartheta(\boldsymbol{X} \wedge X_1, \cdots, X_p) &= \frac{1}{2} \sum_{j=1}^{p} log\ \sigma_{jj} - \frac{1}{2} log\ |\Sigma| \\ &= \frac{1}{2} \sum_{j=1}^{p} log\ \sigma_{jj} - \frac{1}{2} \sum_{\ell=1}^{p} log\ \lambda_\ell,\end{aligned}$$

where $\sigma_{jj} = Var(X_j)$ and λ_ℓ is the ℓth eigenvalue of Σ.

(ii) Let $\boldsymbol{W} = \Gamma \boldsymbol{X}$ be the rotation of the coordinates of X by the matrix Γ of the eigenvectors of Σ. The components of $\boldsymbol{W}$ are uncorrelated and $Var(W_\ell) = \lambda_\ell$. Thus $\vartheta(\boldsymbol{W} \wedge W_1, \cdots, W_p) = 0 \leq \vartheta(\boldsymbol{X} \wedge X_1, \cdots, X_p)$, with equality if and only if X_j's are uncorrelated.

(iii) If $(Y, X_1, \cdots, X_p)$ are jointly normal, then $H[Y|(x_1, \cdots, x_p)]$ is a function of the variances and covariances, and is functionally independent of $(x_1, \cdots, x_p)$. Thus the mutual information is equal to the entropy difference

$$\begin{aligned}\vartheta[Y \wedge (X_1,\cdots,X_p)] &= H(Y) - H[Y|(x_1,\cdots,x_p)] \\ &= -log[1-\rho^2(Y;X_1\cdots X_p)]^{1/2} \\ &= \sum_{j=1}^{p} - log[1-\rho^2(Y;X_j|x_{j-1},\cdots,x_1)]^{1/2},\end{aligned}$$

where $\rho^2(Y;X_1\cdots X_p)$ is the square of the multiple correlation between Y and $X_1,\cdots,X_p$ and $\rho^2(Y;X_j|x_{j-1},\cdots,x_1)$ is the square of the partial correlation between Y and X_j.

The partial mutual information $-log[1-\rho^2(Y;X_j|x_{j-1},\cdots,x_1)]^{1/2}$ gives a measure of relative importance of X_j in regression analysis (Theil 1987, Theil and Chung 1988).

(iv) The normalized index of dependency is $I_C[Y \wedge (X_1,\cdots,X_p)] = \rho^2(Y;X_1\cdots X_p)$.

2.4 Information About A Parameter

Quantification of uncertainty about predicting an outcome of a random draw from a distribution $f(x)$ and comparison of the uncertainties about the outcomes of two probability distributions f_1 and f_2 are of prime interest in many econometrics problems. Examples in regression analysis include comparing the uncertainties associated with: the prior and posterior distributions of the coefficient vector, two posterior distributions of the regression coefficient vector or the sampling distributions of two estimators of the regression coefficient vector based on two different regression matrix structures, etc.

Traditionally, the variance is used for measuring the uncertainty. The widespread use of variance for measuring uncertainty is rooted in the statistical estimation (Fisher 1921). In statistical estimation, Fisher's information is defined as

$$\mathcal{F}(\theta) \equiv \mathcal{F}[f(x|\theta)] = -E_{x|\theta}\left[\frac{\partial^2}{\partial\theta^2} log\, f(x|\theta)\right].$$

$\mathcal{F}(\theta)$ is a measure of information in X, i.e., in $f(x|\theta)$ about the parameter θ, in the sense that $\mathcal{F}(\theta)$ quantifies "the ease with which a parameter can be estimated" by x (Lehmann 1983, p. 120). Inherent in this interpretation is the facts that: (a) X is an unbiased and efficient estimator of θ, so $V(X|\theta) = [\mathcal{F}(\theta)]^{-1}$, and (b) under $f(x|\theta)$, the probabilities are concentrated around the mean value θ.

From the information-theoretic view point, the Fisher information $\mathcal{F}$ is a second order approximation to the discrimination information function $K(f_\theta : f_{\theta+\Delta\theta})$ where θ and $\theta + \Delta\theta$ are two neighboring points in the parameter space and the two distributions f_θ and $f_{\theta+\Delta\theta}$ belong to the same parametric family.

The interpretation of variance as an uncertainty measure about the prediction of an outcome of a random draw from a distribution requires caution. Consider two random variables X and Y with probability distributions f_X and f_Y on the same support. If f_X is flatter than f_Y which assigns high probabilities to the extreme values of Y, then $V(X) < V(Y)$; e.g., $f_X = Beta(1.5, 1.5)$ and $f_Y = Beta(.5, .5)$ are Beta distributions. The outcomes of Y are more volatile, but easier to predict than the outcomes of X. Note that $H(X) > H(Y)$. Ebrahimi and Soofi (1996) showed that for many well-known parametric families of distributions the variance and entropy order similarly in terms of the distributions parameters.

The interpretation of the entropy as a measure of uncertainty about an unknown parameter requires cautions. $H[f(x|\theta)]$ is a measure of uncertainty about an outcome x, not about θ. Sometimes, x is a suitable estimate of θ, e.g., when θ is a location parameter. Ebrahimi and Soofi (1990) interpreted the entropy of the maximum likelihood estimator of a parameter as a measure of information about the parameter being estimated. In such cases, information about x may be interpreted as information about θ. Such indirect uses of entropy as a measure of information should be interpreted accordingly.

In Bayesian statistics involving a parameter θ, the information about the parameter is measured by a discrepancy between the posterior and prior distributions; see, e.g., Goel and DeGroot (1979) and Goel (1983). The difference between the prior entropy $H[\pi(\theta)]$ and the posterior entropy $H[\pi(\theta|x)]$ measures the contribution of data x to the amount of uncertainty about the parameter; see Abel and Singpurwalla (1994) for an interesting example.

The mutual information $\vartheta(\Theta \wedge X)$ provides a measure of expected information in data x about the parameter (Lindley 1956) and has been used in regression problems; see, e.g, Stone (1958), and Soofi (1985, 1990). For $Y = T(X)$, $\vartheta(\Theta \wedge X) \geq \vartheta(\Theta \wedge Y)$, with equality if and only if $T(X)$ is a sufficient statistic for θ. For a fixed $f(x|\theta)$, $\vartheta(\Theta \wedge X)$ is concave in $\pi(\theta)$. However, maximization of $\vartheta(\Theta \wedge X)$ with respect to $\pi(\theta)$ is usually intractable. Lindley (1961) showed that ignorance between two neighboring values θ and $\Delta\theta$ in the parameter space implies that $\vartheta(\Theta \wedge X) \approx 2(\Delta\theta)^2\mathcal{F}(\theta)$, $\mathcal{F}$ being the Fisher information. According to this relationship, Jeffreys prior for θ is an approximate solution to the density that may be obtained by maximization of $\vartheta(\Theta \wedge X)$. Bernardo (1979a, 1979b) developed limiting solution to the maximization of $\vartheta(\Theta \wedge X)$ with respect to $\pi(\theta)$. Hill and Spall (1987) and Spall and Hill (1990) provided approximate solution for the maximization problem.

Zellner (1971) defined an information function for quantifying the information in the data x about a parameter θ with the prior $\pi(\theta)$, which may be written as:

$$\begin{aligned} G[\pi(\theta)] &= E_\pi\{H[\pi(\theta)] - H[f(x|\theta)]\} \\ &= E_\pi\{K[f(x|\theta) : \pi(\theta)]\} \\ &= \vartheta(\Theta \wedge X) + H(\Theta) - H(X). \end{aligned} \tag{2.14}$$

Zellner proposed $G[\pi(\theta)]$ as a criterion function for developing prior distributions that are maximally committed to the data. The prior $\pi^*(\theta)$ that maximizes $G[\pi(\theta)]$ is referred to as the *Maximal Data Information Prior (MDIP).* The first equation in (2.14) is the *a priori* expected information in the data-generating density (likelihood function) which is "purified" from the information in the prior. The second equation in (2.14) shows that $G[\pi(\theta)]$ is the *a priori* expected information for discrimination between the data-generating distribution and the prior. The third equation indicates that $G[\pi(\theta)]$ is a "broader" information criterion for developing prior as compared with $\vartheta(\Theta \wedge X)$. Furthermore, MDIP gives explicit solutions in many problems and is capable of including side information in terms of moment constraints on $\pi(\theta)$; see Zellner (1991) and Soofi (1996) for details.

3 ME Distributions for Regression

In this section, ME distributions for the error terms, coefficients, and precision of a given linear regression are presented. An ME procedure for derivation of regression function is also briefly discussed.

3.1 ME Distributions for Linear Regression

Consider the linear equation:

$$y = X\beta + \varepsilon, \tag{3.1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of observations, X is an $n \times p$ full rank matrix of given regressors, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of error terms.

In order to obtain an ME distribution for the error term for inferential purposes, we need to specify a variation function, $\mathcal{V}(\varepsilon) > 0$ for the error process. The maximum mean value of the variation function υ, signifies the degree of accuracy and its inverse $\varphi = \upsilon^{-1}$, signifies the *precision* of a specified regression in (3.1).

Table 1 gives examples of variation functions and the corresponding ME error distributions obtained using (2.3). As shown in Table 1, the square error variation gives the *normal* distribution and the absolute error variation leads to the *Laplace (double exponential)* error distribution which has a heavier tail than the normal. The logarithmic variation gives the *generalized Student-t* distribution for the error terms (Soofi 1996); the term generalized refers to the fact that the degrees of freedom parameter ν may not be an integer for all precision values $\upsilon(\nu)$. For regression analysis with the Student-t error distribution see Zellner (1976) and Lange, Little, and Taylor (1989).

Table 1. Variation Functions and Maximum Entropy Distributions for Regression Error

Variation Function	$MaxE[\mathcal{V}(\varepsilon)]$	ME Error Distribution
$\mathcal{V}(\varepsilon) = \varepsilon_i^2$ $i = 1, \cdots, n$	σ^2	*Normal* $f^*(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}\varepsilon_i^2}$
$\mathcal{V}(\varepsilon) = \lvert\varepsilon_i\rvert$ $i = 1, \cdots, n$	α	*Laplace* $f^*(\varepsilon_i) = \frac{1}{2\alpha} e^{-\frac{1}{\alpha}\lvert\varepsilon_i\rvert}$
$\mathcal{V}(\varepsilon) = log(\nu + \varepsilon_i^2)$, $\nu \geq 1, \quad i = 1, \cdots, n$	$\psi\left(\frac{\nu}{2} + \frac{1}{2}\right) - \psi\left(\frac{\nu}{2}\right)$	*Generalized t* $f^*(\varepsilon_i) = B(\frac{1}{2}, \frac{\nu}{2})^{-1}\nu^{-1/2}(1 + \frac{\varepsilon_i}{\nu})^{-(\nu+1)/2}$
$\mathcal{V}(\varepsilon) = \varepsilon\varepsilon'$	Σ	*Normal* $f^*(\varepsilon) = (2\pi)^{-n/2}\lvert\Sigma\rvert^{-1/2} e^{-\frac{1}{2}\boldsymbol{\varepsilon}'\lvert\Sigma\rvert^{-1}\boldsymbol{\varepsilon}}$

Notes: $\psi(u) = \Gamma'(u)$, Γ is the gamma function; $B(u, v)$ is the Beta function.

Diagnostics for assessing suitability of a variation function as the description of the error-generating distribution may be developed using the ID distinguishability index (2.6) along the lines of Soofi, Ebrahimi, and Habibullah (1995).

Suppose that we use square error variation

$$E_f(\varepsilon_i^2) \leq \sigma^2, \qquad i = 1, \cdots, n. \tag{3.2}$$

Then the ME gives the following multivariate normal model for the vector of error terms:

$$f^*(\boldsymbol{\varepsilon}; \sigma^2) = N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n), \tag{3.3}$$

where $\boldsymbol{I}_n$ denotes the identity matrix of dimension n.

The independence among the components of $\boldsymbol{\varepsilon}$ in (3.3) is the result of considering solely the marginal variations (3.2) as the information constraints in the ME computation. If there are information available about the interrelationship between the error components, they should be taken into account by formulating appropriate covariation functions (cross-product moments constraints). The ME solution for the error distribution subject to covariation functions and/or nonhomogeneous maximum average variations across the n-dimension will be $N(\mathbf{0}, \Sigma)$ given by (2.4).

Given that $\boldsymbol{\beta}$ and X in (3.1) are not subject to variation, the ME error distribution (3.3) gives the conventional normal regression model

$$f^*(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) = N(X\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n). \tag{3.4}$$

In the ME procedure, the simpler moment assumption (3.2) replaces the more stringent assumption of normality usually made in the traditional regression analysis. But as we have seen, the ME procedure is versatile in producing more general regression models.

Let's now consider variation of $\boldsymbol{\beta}$ in (3.1). Table 2 shows examples of the variation functions for the regression coefficients β_j around the arbitrary constants $m_j, j = 1, \cdots, p$.

If we only incorporate the range of variation of the regression coefficients, then the ME solution is a uniform distribution which is improper when $a = b = \infty$.

For the quadratic variation function $\mathcal{V}(\beta_j) = (\beta_j - m_j)^2$, the information constraints are $E(\beta_j - m_j)^2 \leq \tau^2$, $j = 1, \cdots, p$. These constraints give the ME distribution

$$\pi(\boldsymbol{\beta}; \boldsymbol{m}, \tau^2) = N(\boldsymbol{m}, \tau^2 \boldsymbol{I}_p), \tag{3.5}$$

where $\boldsymbol{m} = (m_1, \cdots, m_p)'$.

The ME distribution (3.5) is the conjugate prior frequently used in Bayesian analysis for $\boldsymbol{\beta}|\sigma^2$ of the likelihood function (3.4).

The classical random effects model is obtained with the combination of (3.4) and (3.5) when $m_j = 0$ for all $j = 1, \cdots, p$. The ME distributions such as (3.5) developed for $\boldsymbol{\beta}$ are also useful for modeling heterogeneity of the regression coefficients among a population of interest which is an important concern in some fields such marketing.

As in the case of error distribution, the lack of incorporating covariation information in the ME computation results in the prior independence among the coefficients. In order to incorporate a covariance structure Ψ as the prior information, then we use $E(\boldsymbol{\beta} - \boldsymbol{m})'(\boldsymbol{\beta} - \boldsymbol{m}) = \Psi$ as the constraints in the ME computation. This constraint gives multivariate normal prior shown in the last row of Table 2.

For example, if wish to use the data covariance structure $\Psi = (X'X)^{-1}$ and $\tau^2 \propto \sigma^{-2}$, then we obtain

$$\pi(\boldsymbol{\beta}; \boldsymbol{m}, \tau^2) = N[\boldsymbol{m}, \tau^2(X'X)^{-1}]. \tag{3.6}$$

which is the g-prior proposed by Zellner (1982).

Table 2. Variation Functions and Maximum Entropy Distributions for Regression Coefficients.

Variation Function	$MaxE[\mathcal{V}(\boldsymbol{\beta})]$	ME Distribution for Coefficient
$\mathcal{V}(\beta_j) = \delta(m_j - a < \beta_j < m_j + b)$ $j = 1, \cdots, p$	1	*Uniform* $f^*(\boldsymbol{\beta}) = (b-a)^{-p}$
$\mathcal{V}(\beta_j) = (\beta_j - m_j)^2$ $j = 1, \cdots, p$	τ^2	*Normal* $f^*(\boldsymbol{\beta}) = (2\pi\tau^2)^{-p/2} e^{-\frac{1}{2\tau^2}(\boldsymbol{\beta}-\boldsymbol{m})'(\boldsymbol{\beta}-\boldsymbol{m})}$
$\mathcal{V}(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{m})(\boldsymbol{\beta} - \boldsymbol{m})'$	$\tau^2\Psi$	*Normal* $f^*(\boldsymbol{\beta}) = (2\pi\tau^2)^{-p/2}\lvert\Psi\rvert^{-1/2} e^{-\frac{1}{2\tau^2}(\boldsymbol{\beta}-\boldsymbol{m})'\Psi^{-1}(\boldsymbol{\beta}-\boldsymbol{m})}$

Note: $\delta(\cdot)$ is the indicator function:

$$\delta(m_j - a < \beta_j < m_j + b) = \begin{cases} 1 & if \;\; \beta_j \in (m_j - a, \; m_j + b), \;\; j = 1, \cdots, p \\ 0 & otherwise. \end{cases}$$

Next I incorporate variation of the precision parameter $\varphi = \sigma^{-2}$ in (3.4). Because φ is positive with probability one, we can consider the types of information constraints shown in Table 3 and obtain the corresponding ME distributions. Except for $log\ \varphi$, the constraints shown in Table 3 may interpreted as variation functions; $log\ \varphi$ may also be interpreted as a variation function if $P(\varphi > 1) = 1$.

Like for the case of the regression coefficients, the ME distributions derived for the precision parameter are useful in the Bayesian and frequentist analysis. The uniform distribution for $log\ \varphi$ is the Jeffreys prior. The first three ME distributions shown in Table 3 are special or limiting cases of the Gamma distribution which is ME using a pair of information constraints. In Bayesian analysis, the Gamma distribution is the conjugate for the normal regression model (3.4). The Gamma distribution is also used in frequentist analyses for modeling is heterogeneity of the regression precision among individuals.

Table 3. Information Constraints and Maximum Entropy Distributions for Regression Precision.

Information Constraint	$MaxE[c(\varphi)]$	ME Distribution for Precision
$c(\varphi) = \delta(a < log\ \varphi < \ b)$	1	*Uniform* $f^*(\varphi) = (b - a)^{-1}$
$c(\varphi) = \varphi$	α	*Exponential* $f^*(\varphi) = \alpha e^{-\frac{\varphi}{\alpha}}$
$c(\varphi) = log\ \varphi, \quad \varphi > a$	$\frac{1}{\alpha}$	*Pareto* $f^*(\varphi) = \alpha a^{\alpha} \varphi^{-\alpha-1}$
$c_1(\varphi) = \varphi$, $c_2(\varphi) = log\ \varphi$	$\alpha\nu$ $\psi(\alpha) - log(\nu)$	*Gamma* $f^*(\varphi) = [\Gamma(\nu)]^{-1}\alpha^{-\nu}\varphi^{\nu-1}e^{-\frac{\varphi}{\alpha}}$

Notes: $\psi(u) = \Gamma'(u)$, Γ is the gamma function. $\delta(\cdot)$ is the indicator function:

$$\delta(a < log\ \varphi < b) = \begin{cases} 1 & if \quad log\ \varphi \in (a,\, b) \\ 0 & otherwise. \end{cases}$$

3.2 ME Regression Functions

A regression function is defined by the conditional expectation and is given by:

$$\begin{aligned} y(\boldsymbol{x}) &\equiv E(Y|\boldsymbol{x}) \\ &= \int y f(y|\boldsymbol{x}) dy = \frac{\int y f(y,\boldsymbol{x}) dy}{\int f(y,\boldsymbol{x}) dy}, \end{aligned}$$

where $\boldsymbol{x} = (x_1, \cdots, x_p)$ is the vector of regressors assumed to be subject to variation. Thus, in principle, one can find the ME joint distribution $f^*(y, \boldsymbol{x})$ that satisfies a set of information moment constraints, and then find the conditional expectation $y(\boldsymbol{x})$.

Ryu (1993) considered the special case when $y(\boldsymbol{x}) > 0$ and noted that $y(\boldsymbol{x})$ is an averaged density with respect to $f(\boldsymbol{x})d\boldsymbol{x}$. Ryu showed that many well-known regression functions can be derived as solutions to

$$\max_{y} \; -\int y(\boldsymbol{x}) log[y(\boldsymbol{x})] f(\boldsymbol{x}) d\boldsymbol{x}$$

subject to constraints

$$\int c_{m_1}(x_1) \cdots c_{m_p}(x_p) y(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = \theta_{m_1, \cdots, m_p}.$$

4 Regression Estimation and Prediction

In this section I discuss information theoretic procedures for estimating regression coefficients and developing forecast distributions.

4.1 MDI Estimation

We have a set of observations $y_1, \cdots, y_n$ generated from an n-variate distribution $f(\boldsymbol{y})$. Our objective is to estimate $\boldsymbol{\theta}$ of the ME distribution $f^*(\boldsymbol{y}; \boldsymbol{\theta})$ implied by the linear relationships (3.1) and the associated variation function $\mathcal{V}(\varepsilon)$. Here $\boldsymbol{\theta}$ denotes the vector of p coefficients and all the parameters related to $\mathcal{V}(\varepsilon)$.

We explicitly differentiate between a convenient mathematical function $f(\boldsymbol{y}; \boldsymbol{\theta})$ termed as *model* which we utilize in practice as an approximation to the unknown true data-generating

$f(y)$. "It may be very likely that the true distribution is in fact too complicated to be represented by a simple mathematical functions such as is given in ordinary textbooks." (Sawa 1978).

Given linear relation (3.1), the variation function $\mathcal{V}(\varepsilon)$, and $\boldsymbol{\theta}$, we derive the ME distribution $f^*(\boldsymbol{y};\boldsymbol{\theta})$ and use it as our *estimate* for the parametric family of the model $f(\boldsymbol{y};\boldsymbol{\theta})$. The symbol f^* underscores the fact that the ME model is being used only as an *approximation* for $f(\boldsymbol{y})$. According to the entropy concentration theorem (Jaynes 1982, van Campenhount and Cover 1981) and the ID distinguishability result of Soofi, et al. (1995), the approximation should be satisfactory if f is a "typical" distribution in the class which f^* is the ME model; i.e, if $ID(f : f^*;\boldsymbol{\theta}) \approx 0$, thus f is not ID distinguishable with f^*.

Therefore, it is natural to estimate the model parameter $\boldsymbol{\theta}$ based on a criterion that improves the model approximation for the data-generating distribution. The MDI or minimum relative entropy loss estimation procedure serves this purpose. [For MDI estimation in other contexts see, e.g., Kullback (1959), James and Stein (1961), Gokhale and Kullback (1978), Haff (1980), Ghosh and Yang (1988), and Soofi and Gokhale (1991a).]

The loss of approximating $f(\boldsymbol{y})$ by an ME model $f^*(\boldsymbol{y};\boldsymbol{\theta})$ with its parameter estimated by $\tilde{\boldsymbol{\theta}}$ is measured

$$2\bar{K}[f(\boldsymbol{y}) : f^*(\boldsymbol{y};\tilde{\boldsymbol{\theta}})] \equiv \frac{2}{n}K[f(\boldsymbol{y}) : f^*(\boldsymbol{y};\tilde{\boldsymbol{\theta}})].$$

The MDI or minimum relative entropy loss estimate $\tilde{\boldsymbol{\theta}}_{MDI}$ of a parameter $\boldsymbol{\theta}$ is defined by:

$$\tilde{\boldsymbol{\theta}}_{MDI} = \arg\min_{\tilde{\boldsymbol{\theta}}} K[f(\boldsymbol{y}) : f^*(\boldsymbol{y};\tilde{\boldsymbol{\theta}})].$$

The Bayesian risk of approximating $f(\boldsymbol{y})$ by an estimated ME model $f^*(\boldsymbol{y};\tilde{\boldsymbol{\theta}})$ is computed by the posterior expectation $2E\{\bar{K}[f(\boldsymbol{y}) : f(\boldsymbol{y};\tilde{\boldsymbol{\theta}})]|\boldsymbol{y}\}$.

The *MDI Bayes (MDIB)* estimate of $\boldsymbol{\theta}$ is defined by

$$\tilde{\boldsymbol{\theta}}_{MDIB} = \arg\min_{\tilde{\boldsymbol{\theta}}} E\{\bar{K}[f(\boldsymbol{y} : f(\boldsymbol{y};\tilde{\boldsymbol{\theta}})]|\boldsymbol{y}\}.$$

The frequentist risk of approximation is found by computing the expected loss with respect to the sampling distribution, $2E_{\tilde{\boldsymbol{\theta}}}\{\bar{K}[f(\boldsymbol{y}) : f(\boldsymbol{y};\tilde{\boldsymbol{\theta}})]\}$.

Decomposing the log-ratio in (2.5) gives

$$2\bar{K}[f(\boldsymbol{y}) : f(\boldsymbol{y};\tilde{\boldsymbol{\theta}})] = 2\bar{H}_f[f^*(\boldsymbol{y};\tilde{\boldsymbol{\theta}})] - 2\bar{H}[f(\boldsymbol{y})], \tag{4.1}$$

where

$$\bar{H}_f[f^*(\boldsymbol{y};\tilde{\boldsymbol{\theta}})] \equiv -\frac{1}{n}E_f[log\ f^*(\boldsymbol{y};\tilde{\boldsymbol{\theta}})]. \tag{4.2}$$

The entropy of the data-generating distribution is free of $\tilde{\boldsymbol{\theta}}$, so $\bar{H}[f(\boldsymbol{y})]$ in (4.1) is sometimes ignored, the loss is measured by expected log-likelihood (4.2), and the MDI estimate may be obtained by

$$\tilde{\boldsymbol{\theta}}_{MDI} = \arg\max_{\tilde{\boldsymbol{\theta}}}\ E_f[log\ f^*(\boldsymbol{y};\tilde{\boldsymbol{\theta}})]. \tag{4.3}$$

Here the cumbersome problem of minimizing the information discrepancy between the unknown data-generating distribution and the ME model is reduced to the simpler problem of maximizing the expected value of a log-likelihood function. But unlike the conventional statistics in which the parameters are often estimated solely by considering a postulated model, the MDI estimation includes both the model and the data-generating distribution. However, at this point the problem is not yet completely solved.

Akaike (1973) proceeded with an MDI parameter estimation by first estimating the expectation in (4.3) using the empirical distribution which assigns a mass of $\frac{1}{n}$ to each data point y_i, $i = 1, \cdots, n$. In this case,

$$\tilde{\boldsymbol{\theta}}_{MDI} = \arg\max_{\tilde{\boldsymbol{\theta}}} \frac{1}{n}\sum_{i=1}^{n} log\ f^*(y_i;\tilde{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}},$$

where $\hat{\boldsymbol{\theta}}$ is the Maximum Likelihood Estimate (MLE) of $\boldsymbol{\theta}$ under the model f^*. Thus from an information theoretic view point, the MLE minimizes an estimated information discrepancy between the data-generating distribution $f(\boldsymbol{y})$ and the ME distribution f^*. Akaike interpreted this approach as an extension of the MLE principle.

Under the assumption that the data-generating distribution is also in the same parametric family as $f^*(\boldsymbol{y};\boldsymbol{\theta})$, Akaike (1973) estimated the information discrepancy for $\hat{\boldsymbol{\theta}}$ by

$$K_A[f^*(\boldsymbol{y},\boldsymbol{\theta}) : f^*(\boldsymbol{y};\hat{\boldsymbol{\theta}})] = \frac{1}{n}log\frac{f^*(\boldsymbol{y};\boldsymbol{\theta})}{f^*(\boldsymbol{y};\hat{\boldsymbol{\theta}})}. \tag{4.4}$$

The consistency of the MLE implies that (4.4) is a consistent estimate of $K[f^*(\boldsymbol{y},\boldsymbol{\theta}) : f^*(\boldsymbol{y};\hat{\boldsymbol{\theta}})]$. Akaike computed an approximate frequentist risk function for the purpose of model selection which will be discussed in Section 5.2.

Consider estimation of the normal ME regression model, $f^*(\boldsymbol{y};\boldsymbol{\beta},\sigma^2) = N(X\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n)$. The normal model is plausible for approximating the data-generating when $f(\boldsymbol{y})$ possesses at least the first two moments, say, $\boldsymbol{\mu}$ and Ω; i.e., $f(\boldsymbol{y}) = f(\boldsymbol{y};\boldsymbol{\mu},\Omega)$. Note that the normal ME regression model uses the specific form $\boldsymbol{\mu} = X\boldsymbol{\beta}$, but the data-generating distribution $f(\boldsymbol{y}) = f(\boldsymbol{y};\boldsymbol{\mu},\Omega)$ is not as restricted.

The loss of approximating $f(\boldsymbol{y};\boldsymbol{\mu},\Omega)$ by the normal regression model with its parameters estimated by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ is given by

$$2\bar{K}[f(\boldsymbol{y};\boldsymbol{\mu},\Omega) : f(\boldsymbol{y};\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)] = 2\bar{H}_f[f^*(\boldsymbol{y};\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)] - 2\bar{H}[f(\boldsymbol{y};\boldsymbol{\mu},\Omega)].$$

For the case of $\Omega = \omega^2\boldsymbol{I}_n$, the expected log-likelihood can be evaluated and the estimation loss is given by

$$\begin{aligned} 2\bar{K}[f(\boldsymbol{y};\boldsymbol{\mu},\omega) : f(\boldsymbol{y};\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)] &= -2\bar{H}[f(\boldsymbol{y};\boldsymbol{\mu},\omega)] + log(2\pi e) \\ &\quad +log\ \tilde{\sigma}^2 + \frac{\omega^2}{\tilde{\sigma}^2} + \frac{(\boldsymbol{\mu} - X\tilde{\boldsymbol{\beta}})'(\boldsymbol{\mu} - X\tilde{\boldsymbol{\beta}})}{n\tilde{\sigma}^2}. \end{aligned} \tag{4.5}$$

Various Bayesian schemes have been suggested for finding the risk of approximating $f(\boldsymbol{y};\boldsymbol{\mu},\omega^2\boldsymbol{I}_n)$ by an estimated normal regression model. Leamer (1979) used the posterior distribution of $(\boldsymbol{\mu},\omega^2)$. Expanding the quadratic form in (4.5) and taking expectation gives

$$\begin{aligned} 2E_{(\boldsymbol{\mu},\omega^2)|\boldsymbol{y}}\ \bar{K}[f(\boldsymbol{y};\boldsymbol{\mu},\omega) : f(\boldsymbol{y};\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)] &= 2\bar{H}[f(\boldsymbol{y};\boldsymbol{\mu},\omega)] + log(2\pi) \\ &+ log\ \tilde{\sigma}^2 + \frac{E(\omega^2|\boldsymbol{y})}{\tilde{\sigma}^2} + \frac{E(\boldsymbol{\mu}'\boldsymbol{\mu})}{n\tilde{\sigma}^2} \\ &+ \frac{1}{n\tilde{\sigma}^2}[\tilde{\boldsymbol{\beta}}'X'X\tilde{\boldsymbol{\beta}} - 2E(\boldsymbol{\mu}'|\boldsymbol{y})X\tilde{\boldsymbol{\beta}}]. \end{aligned} \tag{4.6}$$

The MDIB estimates that minimizes the expected loss (4.6) with respect to $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ are:

$$\tilde{\boldsymbol{\beta}}_{MDIB} = (X'X)^{-1}X'E(\boldsymbol{\mu}|\boldsymbol{y}) \tag{4.7}$$

$$\tilde{\sigma}^2_{MDIB} = E(\omega^2|\boldsymbol{y}) + \frac{1}{n}E(\boldsymbol{\mu}'\boldsymbol{\mu}|\boldsymbol{y}) - \frac{1}{n}E(\boldsymbol{\mu}'|\boldsymbol{y})X(X'X)^{-1}X'E(\boldsymbol{\mu}|\boldsymbol{y}). \tag{4.8}$$

In a given problem these MDIB estimates can be evaluated by using the unrestricted ME distribution $f^*(\boldsymbol{y};\boldsymbol{\mu},\omega^2\boldsymbol{I}_n) = N(\boldsymbol{\mu},\omega^2\boldsymbol{I}_n)$ and some types of ME priors for $(\boldsymbol{\mu},\omega^2)$. When the prior is weak, the MDIB estimates (4.7) and (4.8) are approximately equal to the MLE of $\boldsymbol{\beta}$ and σ^2 under the normal ME model,

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{MDIB} &\approx (X'X)^{-1}X'\boldsymbol{y} \equiv \boldsymbol{b} \\ \tilde{\sigma}^2_{MDIB} &\approx \frac{1}{n}(\boldsymbol{y}-X\tilde{\boldsymbol{b}})'(\boldsymbol{y}-X\tilde{\boldsymbol{b}}) \equiv \hat{\sigma}^2 \end{aligned} \tag{4.9}$$

Sawa (1978) assumed that $f(\boldsymbol{y}) = N(\boldsymbol{\mu},\omega^2\boldsymbol{I}_n)$, and defined a risk function in terms of the posterior distributions of the parameters of the normal regression model $f^*(\boldsymbol{y};\boldsymbol{\beta},\sigma^2)$. Using diffuse priors $\boldsymbol{\beta}$ and σ^2), he found that

$$2E_{(\boldsymbol{\beta},\sigma^2)|\boldsymbol{y}}\left\{\bar{H}_f[f^*(\boldsymbol{y};\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)]\right\} \geq log(2\pi) + log\;\tilde{\sigma}^2 + \frac{n-p}{n-p-2}\left(1+\frac{p}{n}\right)\frac{s^2}{\tilde{\sigma}^2}, \tag{4.10}$$

where, s^2 is the mean square error of the least square regression.

Young (1987) defined a risk function as $E_{(\boldsymbol{\mu},\omega^2)|\boldsymbol{y}}E_{(\boldsymbol{\beta},\sigma^2)|\boldsymbol{y}}\;\bar{K}[f(\boldsymbol{y};\boldsymbol{\mu},\omega) : f(\boldsymbol{y};\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)]$; see Section 5.2.

Sawa (1978) also found an approximation for the frequentist risk of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$. He showed that if the components of $\boldsymbol{y}$ are symmetrically distributed with the same kurtosis as the normal distribution, then the frequentist risk of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ is approximately,

$$\begin{aligned} 2E_{\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2}\;\bar{K}[f(\boldsymbol{y};\boldsymbol{\mu},\omega) : f(\boldsymbol{y};\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)] &= 2\bar{H}[f(\boldsymbol{y};\boldsymbol{\mu},\omega)] + log(2\pi e) \\ &\quad + log\;\sigma_0^2 + \frac{p_k+1}{n}\left(\frac{\omega^2}{\sigma_0^2}\right) - \frac{1}{n}\left(\frac{\omega^2}{\sigma_0^2}\right)^2. \end{aligned} \tag{4.11}$$

The quantity σ_0^2 is defined below.

The solutions to the minimization of $K[f(\boldsymbol{y};\boldsymbol{\mu},\Omega) : f^*(\boldsymbol{y};\boldsymbol{\beta},\sigma^2)]$ with respect to the model parameters $\boldsymbol{\beta}$ and σ^2 are:

$$\begin{aligned} \boldsymbol{\beta}_0 &= (X'X)^{-1}X'\boldsymbol{\mu} \\ \sigma_0^2 &= \frac{1}{n}\boldsymbol{\mu}'[\boldsymbol{I}_n - X(X'X)^{-1}X']\boldsymbol{\mu} + \omega^2. \end{aligned}$$

These quantities are referred to as *pseudo-true parameter values.* The MLE of $\boldsymbol{\beta}$ is unbiased for the pseudo-true parameter value $\boldsymbol{\beta}_0$, and the MLE of σ^2 is asymptotically unbiased for the pseudo-true parameter value σ_0^2, Sawa (1978).

4.2 MDI Method of Moments

The *MDI moment (MDIM)* estimate of $\boldsymbol{\theta}$ is defined by the solution to the following constrained relative entropy loss problem:

$$\min_{\boldsymbol{\theta}} K[f(\boldsymbol{y}) : f^*(\boldsymbol{y}; \boldsymbol{\theta})]$$

subject to

$$\int T_m(\boldsymbol{y}) f(\boldsymbol{y}) d\boldsymbol{y} = \bar{T}_j(\boldsymbol{y}), \quad m = 1, \cdots, M,$$

where $\bar{T}_j(\boldsymbol{y})$ is a sample moment of interest.

As an specific example, we construct MDIM estimates for the parameters of the normal regression model $f^*(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) = N(X\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$.

Suppose that our data consist of

$$y_{1k_1}, \cdots, y_{nk_n}, \quad k_i = 1, \cdots, n_i \geq 1, \quad i = 1, \cdots, n.$$

The MDIM estimates of $\boldsymbol{\beta}$ and σ^2 are found by the solutions of:

$$\min_{\boldsymbol{\beta}, \sigma^2} K[f(\boldsymbol{y}) : f^*(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2)] \tag{4.12}$$

subject to

$$\int y_i f(\boldsymbol{y}) d\boldsymbol{y} = \bar{y}_i, \quad i = 1, \cdots, n, \tag{4.13}$$

$$\int (y_i - \bar{y}_i)^2 f(\boldsymbol{y}) d\boldsymbol{y} = s_i^2, \quad i = 1, \cdots, n, \tag{4.14}$$

where

$$\bar{y}_i = \frac{1}{n_i} \sum_{k_i=1}^{n_i} y_{ik_i}$$

$$s_i^2 = \frac{1}{n_i} \sum_{k_i=1}^{n_i} (y_{ik_i} - \bar{y}_i)^2.$$

Assuming $f(\boldsymbol{y})$ satisfies the regularity conditions required for taking the derivative to inside the integral sign, the MDIM estimates are found as:

$$\tilde{\boldsymbol{\beta}}_{MDIM} = (X'X)^{-1} X' \bar{\boldsymbol{y}} \tag{4.15}$$

$$\tilde{\sigma}^2_{MDIM} = \frac{1}{n} \bar{\boldsymbol{y}}' [\boldsymbol{I}_n - X(X'X)^{-1} X'] \bar{\boldsymbol{y}} + \frac{1}{n} \sum_{i=1}^{n} s_i^2, \tag{4.16}$$

where $\bar{\boldsymbol{y}} = (\bar{y}_1, \cdots, \bar{y}_n)'$. The MDI estimate (4.15) was introduced in Soofi (1985).

In the MDIM procedure, the unknown values μ_i are estimated by $\bar{y}_i$ using constraint (4.13). This is in line with the common practice of using the regression estimate of $\hat{y}_i$ as the point estimate of the conditional expected value corresponding to $\boldsymbol{x}_i$. Then, the mean variation function in each dimension is estimated by s_i^2 using constraint (4.14). Finally, the information discrepancy between the unknown data-generating distribution and the ME regression model is minimized.

The results obtained using the MDIM procedure is akin to those obtained using the conventional techniques. The two components in the MDIM estimate of the error variance are related to the well-known quantities in regression analysis. The first term in (4.16) is the component of variance due to *lack of fit* of the regression (3.1) to the data and the second term is the component of variance due to *pure error.*

For the case of a single observation per dimension, $n_i = 1,\ \ \bar{y}_i = y_i,\ \ s_i^2 = 0$, the MDIM estimates (4.15) and (4.16) are equivalent to the usual MLE of $\boldsymbol{\beta}$ and σ^2. Thus, the MDIM estimates of the parameters of the normal regression model possess all the properties of the MLE.

The relationship between the MDIM and MLE is similar to a duality that exists between the MLE and the Internal Constraint Problem (ICP) formulation of Gokhale and Kullback (1978), an estimation method extensively used in the information-theoretic analysis of contingency tables. In the ICP formulation, the discrimination information function between an unknown distribution f and a known distribution g, $K(f : g)$, is minimized with respect to f subject to constraints (2.3) with the information moment values θ_j obtained from the data. When g is uniform, then the MLE of f^* and the MDI estimate of f are equivalent. The above MDIM procedure is similar to ICP in that the constraints use the data moments, but in (4.12), the reference distribution is not completely known and is not uniform. When the reference distribution is not uniform, the equivalence with MLE is problem-specific, thus may not always hold.

The MDIM procedure is also in line with the approach of Sawa (1978). In the MDIM procedure, $\boldsymbol{\beta}$ and σ^2 are estimated directly, instead of first developing the MDI model with the pseudo-true parameters and then estimating them by the MLE.

The usual frequentist inference can be done using the sampling properties of the MDIM estimates (4.15) and (4.16) under the normal ME model $f^*(\boldsymbol{y};\boldsymbol{\beta},\sigma^2) = N(X\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n)$. The usual Bayesian inference can be done using the normal ME likelihood function, selecting a prior for $\boldsymbol{\beta}$ from the ME distributions in Table 2, and selecting a prior for the precision parameter from the ME distribution shown in Table 3.

Application of MDIM to other ME error distributions such as those shown in Table 1 will lead to new regression analyses. The use of other ME error distributions as likelihood functions, and other ME priors will lead to new Bayesian regression analysis.

4.3 An MDI Predictive Density

Let $\boldsymbol{y}_N[X]$ be an $m\times 1$ vector of forecasts corresponding to the m new vectors of explanatory variables arranged in the rows of X_m. The normal ME regression model (3.4) implies that the ante-data forecast distribution

$$f^*(\boldsymbol{y}_N[X_m];\boldsymbol{\beta},\sigma^2) = N(X_m\boldsymbol{\beta},\sigma^2\boldsymbol{I}_m). \tag{4.17}$$

Since the parameters are unknown, the ante-data forecast distribution (4.17) is not usable. Several frequentist and Bayesian procedures for developing predictive distributions free of unknown parameters are available, see Geisser (1993).

Many of the known Bayesian and frequentist predictive distributions are in the class

$$\mathcal{G} = \left\{ g(\boldsymbol{y}_N[X_m]|D) : g(\boldsymbol{y}_N[X_m]|D) = h\left(\frac{\boldsymbol{y}_N[X_m] - X_m\boldsymbol{b}}{\sqrt{n\hat{\sigma}^2}}\right)\right\},$$

where $g(\cdot)$ is a density and $h(\cdot)$ is a scaler function and D refers to the observed data $(X,\boldsymbol{y})$; Levy and Perng (1986) and Keyes and Levy (1996).

Levy and Perng (1986) considered the following minimization of the expected discrimination information function:

$$\min_{g\in\mathcal{G}} E\boldsymbol{y}\{K(f^*(\boldsymbol{y}_N[X_m]) : g(\boldsymbol{y}_N[X_m]|\boldsymbol{y}))\}, \tag{4.18}$$

where the expectation is with respect to the normal ME model $f^*(\boldsymbol{y};\boldsymbol{\beta},\sigma^2) = N(X\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n)$ and $f^*(\boldsymbol{y}_N[X_m])$ is the ante-data forecast distribution (4.17). The solution is the m-dimensional

Student-t distribution with $n-p$ degrees of freedom,

$$g^* = t\left(n-p,\ X_m\boldsymbol{b},\ n\hat{\sigma}^2[\boldsymbol{I}_m + X_m(X'X)^{-1}X'_m]\right), \tag{4.19}$$

where $X_m\boldsymbol{b}$ is the location parameter and $\hat{\sigma}^2[\boldsymbol{I}_m + X_m(X'X)^{-1}X'_m]$ is the dispersion matrix. Keyes and Levy (1996) extended this result to multivariate linear models.

The predictive distribution (4.19) is the one obtained in Bayesian regression based on the Jeffreys prior $\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$ and the normal ME likelihood. This coincidence is due to the fact that in (4.18) the objective is to find the predictive density which, on average, has the least information discrepancy with the ante-data ME density $f^*(\boldsymbol{y}_N[X_m])$. That is, in (4.18), we search for the member of $\mathcal{G}$ which is closet to the least informative ante-data density and we find the one based on the Jeffreys non-informative prior as the solution.

4.4 Bayesian Method of Moments

The Bayesian Method of Moments (BMOM), recently proposed by Zellner (1994), combines the use of sample moments and ME procedure. The BMOM combines the least square with the ME procedure and produces posterior (conditional on data) results without a need for introducing likelihood functions and prior densities; i.e., the BMOM bypasses the Bayes Theorem.

Zellner considered the linear equation (3.1) in which $\boldsymbol{y}$, X and $\boldsymbol{\beta}$ are defined as before, and $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}$ is the vector of the realized error terms. The data $D = (\boldsymbol{y}, X)$ is given, thus the quantities X and $\boldsymbol{y}$ are not subject to variation. But the quantities $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are unknown and subject to variation.

The posterior means of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are obtained based on the first moment assumption of BMOM:

$$X'E(\boldsymbol{u}|D) = 0. \tag{4.20}$$

Note that if (3.1) includes an intercept term, then $E(\bar{u}|D) = n^{-1}\sum_{i=1}^{n} E(u_i|D) = 0$. Taking the expectation of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ in (3.1) gives

$$E(\boldsymbol{\beta}|D) = (X'X)^{-1}X'\boldsymbol{y} = \boldsymbol{b} \tag{4.21}$$

$$E(\boldsymbol{u}|D) = \boldsymbol{y} - X\boldsymbol{b} = \hat{\boldsymbol{u}}, \tag{4.22}$$

where $\hat{\boldsymbol{u}}$ is the vector of least square residuals.

The use of (4.21), (4.22), and (4.20) gives

$$\begin{aligned} Var(\boldsymbol{\beta}|D) &= E(\boldsymbol{\beta}-\boldsymbol{b})(\boldsymbol{\beta}-\boldsymbol{b})' \\ &= (X'X)^{-1}X'E(\boldsymbol{u}-\hat{\boldsymbol{u}})(\boldsymbol{u}-\hat{\boldsymbol{u}})'X(X'X), \end{aligned} \tag{4.23}$$

where the covariance structure of $\boldsymbol{u}$ is a solution to the functional equation,

$$E(\boldsymbol{u}-\hat{\boldsymbol{u}})(\boldsymbol{u}-\hat{\boldsymbol{u}})' = X(X'X)^{-1}X'E(\boldsymbol{u}-\hat{\boldsymbol{u}})(\boldsymbol{u}-\hat{\boldsymbol{u}})'X(X'X)^{-1}X'.$$

Zellner proposed the following solution to the functional equation as the second moment assumption of BMOM:

$$Var(\boldsymbol{u}|D,\sigma^2) = E[(\boldsymbol{u}-\hat{\boldsymbol{u}})(\boldsymbol{u}-\hat{\boldsymbol{u}})'|\sigma^2] = X(X'X)X'\sigma^2. \tag{4.24}$$

Using (4.24) in (4.23) gives the posterior covariance matrix of $\boldsymbol{\beta}$ conditional on σ^2 as

$$Var(\boldsymbol{\beta}|D,\sigma^2) = (X'X)^{-1}\sigma^2.$$

When (3.1) includes an intercept term, some algebraic manipulations of (4.24) gives the posterior expectation of σ^2 as the mean square error of the least square regression,

$$E(\sigma^2|D) = \frac{(\boldsymbol{y}-X\boldsymbol{b})'(\boldsymbol{y}-X\boldsymbol{b})}{n-p} \equiv s^2.$$

The forecast for a new vector $\boldsymbol{x}_N$ is given by $y_N[\boldsymbol{x}_m] = \boldsymbol{x}_m'\boldsymbol{\beta} + u_m$. Thus as a function of $\boldsymbol{\beta}$ and u_m, the forecast is subject to variation. Conditional on σ^2, the mean and variance of the forecast are:

$$\begin{aligned} E(y_N[\boldsymbol{x}_m])|D,\sigma^2) &= \boldsymbol{x}_n'\boldsymbol{b} \\ Var(y_N[\boldsymbol{x}_m])|D,\sigma^2) &= [1+\boldsymbol{x}_m'(X'X)^{-1}\boldsymbol{x}_m]\sigma^2. \end{aligned}$$

Posterior and predictive distributions of various quantities of interest for BMOM regression analysis are shown in Table 4. The normal and exponential distributions are obtained directly by the ME procedure based on the BMOM derived in terms of the data. The Laplace (Double Exponential) distributions are derived by integrating out σ^2 from the joint density

Table 4. Maximum Entropy Posterior and Predictive Distributions for BMOM Regression.

Quantity	ME Distribution	Mean	Variance
$\boldsymbol{\beta} \mid D,\ \sigma^2$	*Normal*	$\boldsymbol{b}$	$(X'X)^{-1}\sigma^2$
$u_i \mid D,\ \sigma^2$	*Normal*	$\hat{u}_i$	$\boldsymbol{x}_i'(X'X)^{-1}\boldsymbol{x}_i\sigma^2$
$Y_N[\boldsymbol{x}_m] \mid D,\ \sigma^2$	*Normal*	$\boldsymbol{x}_m'\boldsymbol{b}$	$[1+\boldsymbol{x}_m'(X'X)^{-1}\boldsymbol{x}_m]\sigma^2$
$\sigma^2 \mid D$	*Exponential*	s^2	s^4
$\beta_j \mid D$	*Laplace*	b_j	$s_j^2 = (j,j)\underline{th}\ element\ of\ (X'X)^{-1}s^2$
$\boldsymbol{\ell}'\boldsymbol{\beta} \mid D$	*Laplace*	$\boldsymbol{\ell}'\boldsymbol{b}$	$\boldsymbol{\ell}'(X'X)^{-1}\boldsymbol{\ell}s^2,\ \ \boldsymbol{\ell}' = (\ell_1,\cdots,\ell_p)$
$u_i \mid D$	*Laplace*	$\hat{u}_i$	$\boldsymbol{x}_i'(X'X)^{-1}\boldsymbol{x}_i s^2$
$Y_N[\boldsymbol{x}_m] \mid D$	*Laplace*	$\boldsymbol{x}_m'\boldsymbol{b}$	$[1+\boldsymbol{x}_m'(X'X)^{-1}\boldsymbol{x}_m]s^2$

Notes: D refers to the data $(X, \boldsymbol{y})$. The density of the Laplace (Double Exponential) distribution with mean ν and variance ω^2 is:

$$f(z|\nu,\omega) = \frac{1}{\sqrt{2}\omega}e^{-\frac{\sqrt{2}}{\omega}|z-\nu|}.$$

given by the product of the respective normal conditional density and the exponential density for $\sigma^2|D$.

The Laplace predictive distribution, derived based on BMOM, generally gives wider intervals than those obtained using the normal and the Student-t predictive distributions found in the conventional Bayesian and frequentist regression approaches.

4.5 ME Estimation With Undersize Sample

An undersize sample refers to the situation when $n < p$ in the linear relationship (3.1). In the science and engineering fields related to image reconstruction, the problem is referred to as ill-posed inverse problem and ME inversion method is available to solve the problem (Gull and Daniell 1978, Skilling and Bryan 1984, Gull 1989). In the ME inversion method, the *image* $\boldsymbol{\beta}$ is a high dimensional vector of positive elements that are reconstructed based on the noisy data vector $\boldsymbol{y}$ which has a dimension much lower than the rank of the linear operator X; i.e., $n \ll p$.

The ME inversion method solves the ill-posed problem using the following formulation:

$$\max_{\boldsymbol{\beta}} - \sum_{j=1}^{p} \beta_j log(\beta_j) \quad subject \ \ to \quad (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}) \leq \theta. \tag{4.25}$$

The ME estimate is given by

$$\boldsymbol{b}_{ME}(\eta) = \arg\min_{\boldsymbol{\beta}} \ \sum_{j=1}^{p} \beta_j log(\beta_j) + \eta_1(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}). \tag{4.26}$$

The solution is found using an optimization routine. Skilling and Bryan (1984) have developed special routine for the ME inversion method.

Strictly speaking, $\sum_{j=1}^{p} \beta_j log(\beta_j)$ in (4.25) is not a bona fide entropy because the normalizing constraint $\sum_{j=1}^{p} \beta_j = 1$ is not included. But this causes no problem since the solution of (4.26) are positive and can be normalized if so desired.

In the traditional terms, the solution to the dual of (4.25) is the following constrained (regularized) least square estimate

$$\boldsymbol{b}_{ME}(\eta) = \boldsymbol{b}_{LS}(\eta) = \arg\min_{\boldsymbol{\beta}} \ (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}) + \eta_2 \sum_{j=1}^{p} \beta_j log(\beta_j).$$

The solution depends on the parameter $\eta = \eta(\theta)$ which may be chosen based on some statistical criteria such as cross-validation; for more detail see Donoho et al (1992) and discussions following that article.

Next I discuss two ME estimation methods, proposed by Theil and Laitinen (1980) and Vinod (1982), that avoid singularity of $X'X$ when $n < p$. These methods are based on viewing the rows of X in (3.1) as samples from a p-dimensional random variable $\boldsymbol{x}$.

Under the assumptions that $\boldsymbol{\beta}$ is not subject to variation, X is subject to variation, and $E(X'\boldsymbol{\varepsilon}) = \mathbf{0}$, the regression coefficient is given by

$$\boldsymbol{\beta} = \Sigma_x^{-1} \boldsymbol{\sigma}_{yx},$$

where $\Sigma_x = [\sigma_{jk}]$ is the covariance matrix of the explanatory variables $X_1, \cdots, X_p$ and $\boldsymbol{\sigma}_{yx} = (\sigma_{y,1}, \cdots, \sigma_{y,p})'$ is the vector of the covariances of y with X_k, $k = 1, \cdots, p$.

In the case of random regressors, $X'X$ and $X'y$ are estimates of the cross-product moments $E(\boldsymbol{x}\boldsymbol{x}')$ and $E(\boldsymbol{x}y)$ obtained by the sample second order moments.

Theil and Laitinen (1980) proposed estimating Σ_x by the second order moments of the ME distribution that they developed for $\boldsymbol{x}$ by assuming that $F(\boldsymbol{x})$ is continuous. Let $x_j^1 < x_j^2 < \cdots < x_j^n$ denote the order statistics for the sample $x_{j1}, \cdots, x_{jn}$, and let the intermediate points ξ_j^i be defined as $x_j^i < \xi_j^i < x_j^{i+1}$, $i = 1, \cdots, n-1$. The set of all intermediate points

$$\xi_j^1 < \xi_j^2 < \cdots < \xi_j^{n-1}, \quad j = 1, \cdots p,$$

partitions $\Re^p$ into n^p regions. The partitions are either bounded hyper-rectangular regions with sides given by the interval segments connecting the pairs $(\xi_j^i,\ \xi_j^{i-1})$, $i = 1, \cdots, n-1$, or semi-bounded hyper-rectangular regions that have open-ended intervals of types $(-\infty,\ \xi_j^1]$ and/or $[\xi_j^{n-1}, \infty)$ for a number of their sides. There are n regions $R_1, \cdots, R_n$ each containing one data point $\boldsymbol{x}_i$ and all other regions are empty. Theil and Laitinen constructed the ME(ξ) distribution of $\boldsymbol{x}$ subject to the *mass-preserving* constraint

$$\int_{R_i} f(\boldsymbol{x}) d\boldsymbol{x} = \frac{1}{n}, \quad i = 1, \cdots, n, \tag{4.27}$$

and the *mean-preserving* constraint

$$\int x_j f(\boldsymbol{x}) d\boldsymbol{x} = \bar{x}_j \quad j = 1, \cdots, p. \tag{4.28}$$

The constraints (4.27) and (4.28) produce a p-variate ME(ξ) distribution F^* with the following properties:

(i) The ME(ξ) density, $f^*(\boldsymbol{x}) > 0$ if and only if $\boldsymbol{x} \in R_i$ for an $i = 1, \cdots, n$.

(ii) The ME(ξ) distribution, $F^*(\boldsymbol{x})$ is the product of piecewise uniform marginals when $\boldsymbol{x} \in R_i$, R_i is a bounded hyper-rectangular region. $F^*(\boldsymbol{x})$ is the product of uniform and exponential marginals when $\boldsymbol{x} \in R_i$, R_i is a semi-bounded hyper-rectangular region.

(iii) The mean-preserving constraint makes the intermediate points ξ_j^i to be the *primary midpoints* $\xi_j^i = \bar{x}_j^i = \frac{1}{2}(x_j^i + x_j^{i+1})$, $i = 1, \cdots, n-1$. The mean of the individual intervals are given by the *secondary midpoints*

$$E(\boldsymbol{X}_j | \boldsymbol{X} \in R_i) = \bar{\bar{x}}_j^i = \frac{1}{2}(\xi_j^{i-1} + \xi_j^i) = \begin{cases} \frac{3}{4}x_j^1 + \frac{1}{4}x_j^2 & \mathit{for}\ i = 1 \\ \frac{1}{4}x_j^{i-1} + \frac{1}{2}x_j^i + \frac{1}{4}x_j^{i+1} & \mathit{for}\ i = 2, \cdots, n-1 \\ \frac{1}{4}x_j^{n-1} + \frac{3}{4}x_j^n & \mathit{for}\ i = n, \end{cases}$$

where $\xi_j^0 = x_j^1$ and $\xi_j^n = x_j^n$. Thus the sample mean $\bar{x}_j$ is also the mean of the secondary midpoints, $\bar{\bar{x}}_j^1, \cdots, \bar{\bar{x}}_j^n$.

(iv) The covariance matrix of $f^*(\boldsymbol{x})$ is S^* with elements defined by

$$\begin{aligned} s_{jk}^* &= \frac{1}{n}\sum_{i=1}^{n}(\bar{\bar{x}}_{ji} - \bar{x}_j)(\bar{\bar{x}}_{ki} - \bar{x}_k) && (4.29) \\ &+ \frac{\delta_{ij}}{12n}\left[\sum_{i=1}^{n}(\xi_j^i - \xi_j^{i-1})^2 + 2(\xi_j^1 - \xi_j^0)^2 + 2(\xi_j^n - \xi_j^{n-1})^2\right], && (4.30) \end{aligned}$$

where $\bar{\bar{x}}_{ji}$ are the secondary midpoints rearranged according to the original data index and δ_{ij} is the Kronecker delta. The ME(ξ) variance s_{jj}^* is smaller than the sample variance and the amount of shrinkage is given by

$$\begin{aligned} s_{jj}^* &= \frac{1}{n}\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2 \\ &- \frac{1}{24n}\left[6\sum_{i=1}^{n-1}(x_j^{i+1} - x_j^i)^2 + \sum_{i=2}^{n-1}(x_j^{i+1} - x_j^{i-1})^2\right]. \end{aligned}$$

The ME(ξ) covariance matrix may be written as $S^* = \bar{\bar{S}} + D_\xi$, where $\bar{\bar{S}}$ is the covariance matrix of the secondary midpoints whose elements are given by the expression (4.29) and D_ξ denotes the diagonal matrix with elements shown in the expression (4.30). The ME(ξ) covariance S^* is positive definite.

The ME(ξ) estimate of the linear regression coefficient is

$$\boldsymbol{b}_{ME}(\xi) = S^{*-1}\boldsymbol{s}_{y\boldsymbol{x}}^* = (\bar{\bar{S}} + D_\xi)^{-1}\bar{\bar{\boldsymbol{s}}}_{y\boldsymbol{x}}, \tag{4.31}$$

where $\boldsymbol{s}_{y\boldsymbol{x}}^*$ is the vector of ME(ξ) covariances between y and the explanatory variables computed using the secondary midpoints $\bar{\bar{y}}_i$ and $\bar{\bar{x}}_{ji}$. The last expression in (4.31) shows that $\boldsymbol{b}_{ME}(\xi)$ may be computed as a ridge estimate of the secondary midpoints. The ridge values are given by the secondary midpoints as shown in (4.30). Because of this ridge structure, $\boldsymbol{b}_{ME}(\xi)$ can be computed for undersize samples.

Meisner (1980) compared the risks of $\boldsymbol{b}_{ME}(\xi)$ and the ordinary least square estimate $\boldsymbol{b}$ under the quadratic loss when the data is generated from a multivariate normal distribution. Because of complications, he only considered $n = 2$ and $p = 2, 3$. For $p = 2$, $\boldsymbol{b}_{ME}(\xi)$ compares favorably with $\boldsymbol{b}$ over most of the parameter space. For $p = 3$, $\boldsymbol{b}$ does not exist.

Vinod (1982) developed ME(d) estimate for linear regression coefficient using the above procedure but relaxing the assumption of continuity for $F(\boldsymbol{x})$. He formulated the problem in terms of the data being subject to rounding errors of magnitudes $d_j \geq 0$. Thus $F(x_j)$ is locally continuous in the neighborhood V_i of the data points defined by $V_i = R_{1i} \times \cdots \times R_{pi}$ where

$$R_{ji} = [x_{ji} - d_j,\ x_{ji} + d_j], \quad d_j \geq 0, \quad j = 1, \cdots, p, \quad i = 1, \cdots, n.$$

The mass preserving constraint is given by

$$\int_{V_i} f(\boldsymbol{x}) dF(\boldsymbol{x}) = \frac{1}{n}, \quad i = 1, \cdots, n. \tag{4.32}$$

The ME(d) distribution subject to the mass-preserving constraint (4.32) and the mean-preserving constraint (4.28) is the product of uniform marginals if $\boldsymbol{x} \in V_i$. For $d_j = 0,\ F^*(x_j)$ is the usual empirical distribution.

The covariance matrix S^{**} of the ME(d) has much simpler structure than that shown in (4.29) and (4.30) for S^*. The elements of S^{**} are given by

$$s_{jk}^{**} = \frac{1}{n} \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) + \frac{\delta_{jk} d_j d_k}{3},$$

where δ_{jk} is the Kronecker delta.

The ME(d) covariance matrix may be written as $S^{**} = S_x + D_d$ where S_x is the matrix of the usual sample second order central moments s_{jk}, and D_d is the diagonal matrix with elements d_j^2 in the diagonal. Thus, S^{**} is positive definite. Note that $s_{jk}^{**} = s_{jk},\ j \neq k$.

The ME(d) estimate of the linear regression coefficient is

$$\boldsymbol{b}_{ME}(d) = S^{**-1} \boldsymbol{s}_{y\boldsymbol{x}} = (S_x + D_d)^{-1} \boldsymbol{s}_{y\boldsymbol{x}},$$

where $\boldsymbol{s}_{y\boldsymbol{x}}$ is the vector of usual second order central moments between y and the explanatory variables. Thus, $\boldsymbol{s}_{y\boldsymbol{x}}$ also has a ridge structure and may be computed for undersize samples. The ridge values are given by the magnitudes of the rounding errors. When the variables are not subject to measurement error, $\boldsymbol{b}_{ME}(d)$ reduces to the ordinary least square estimate.

5 Discriminating Between Alternative Models

In this section I consider the problem of discriminating between alternative linear relationships

$$M_k : \boldsymbol{y} = X_k \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k \quad k = 1, 2, \cdots \tag{5.1}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}_k$ are the $n \times 1$ vectors, X_k is $n \times p_k$ full rank matrix, and $\boldsymbol{\beta}_k$ is $p_k \times 1$.

The issue of regression models being nested or non-nested often arises in discussions of discriminating between alternative models. Pesaran (1987) operationalized the concept of nested and non-nested hypotheses in terms of the discrimination information. As an example, Pesaran discussed the issue for regression models using the usual normality assumption and the concept of "true" parameter. I adapt Pesaran's approach and discuss the issue along the lines of the ME and MDI developments of the previous sections.

Let $\mathcal{V}_k(\boldsymbol{\varepsilon}_k)$ be the variation function for the error, φ_k be the corresponding precision parameter, and $f_k^*(\boldsymbol{\varepsilon}_k)$ be the implied ME model for the error term in M_k. If $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k', \varphi_k)$ and X are not subject to variation, then each ME model for the error term in (5.1) implies an ME distribution $f_k^*(\boldsymbol{y}; X_k, \boldsymbol{\theta}_k)$ for $\boldsymbol{y}$ under M_k.

A model M_k is said to be *nested* in the model M_ℓ, denoted by $M_k \preceq M_\ell$, if and only if

$$\bar{K}(\boldsymbol{\theta}_{k0}, \boldsymbol{\theta}_\ell^*; X_k, X_\ell) \equiv \frac{1}{n} \min_{\boldsymbol{\theta}_\ell \in \Theta_\ell} K[f_k^*(\boldsymbol{y}; X_k, \boldsymbol{\theta}_{k0}) : f_\ell^*(\boldsymbol{y}; X_\ell, \boldsymbol{\theta}_\ell)] = 0 \tag{5.2}$$

for all admissible pseudo-true parameter values $\boldsymbol{\theta}_{k0}$ in the parameter space Θ_k of M_k. If (5.2) holds for some but not all admissible pseudo-true parameter values, then M_k is said to be *partially non-nested* with respect to M_ℓ. If (5.2) does not hold for any admissible pseudo-true parameter value, then the M_k is said to be *globally non-nested* with respect to M_ℓ. If $M_k \preceq M_\ell$ and $M_\ell \preceq M_k$, then the two models are said to be *observationally equivalent.*

If we use quadratic variation function and $\varphi_k = \sigma_k^{-2}$, then

$$M_k : f_k^*(\boldsymbol{y}; X_k, \boldsymbol{\beta}_k, \sigma_k^2) = N(X_k \boldsymbol{\beta}_k, \sigma_k^2 \boldsymbol{I}_n). \tag{5.3}$$

Using (2.9) with $\boldsymbol{\mu}_f = X_k \boldsymbol{\beta}_k$ and $\boldsymbol{\mu}_g = X_\ell \boldsymbol{\beta}_\ell$, and minimizing with respect to $(\boldsymbol{\beta}_\ell', \sigma_\ell^2)$ gives

$$\begin{aligned} \boldsymbol{\beta}_\ell^* &= (X_\ell' X_\ell)^{-1} X_\ell' X_k \boldsymbol{\beta}_{k0} \\ \sigma_\ell^{2*} &= \sigma_{k0}^2 + (X_k \boldsymbol{\beta}_{k0})' \left[\boldsymbol{I}_n - X_\ell (X_\ell' X_\ell)^{-1} X_\ell'\right] X_k \boldsymbol{\beta}_{k0}. \end{aligned} \tag{5.4}$$

These MDI parameters give

$$\bar{K}(\beta_{k0}, \sigma^2_{k0}, \beta^*_\ell, \sigma^{2*}_\ell; X_k, X_\ell) = \frac{1}{2} log \left(\frac{\sigma^{2*}_\ell}{\sigma^2_{k0}} \right).$$

Therefore, $M_k \preceq M_\ell$ if and only if $\sigma^{2*}_\ell = \sigma^2_{k0}$. That is, the second term in (5.4) is identically equal to zero, which holds when $\beta_{k0} = \mathbf{0}$ or when $[I_n - X_\ell(X'_\ell X_\ell)^{-1} X'_\ell] X_k = \mathbf{0}$. The first case is trivial. The second condition implies that $M_k \preceq M_\ell$ if X_k is in the linear space spanned by the columns of X_ℓ. The usual case of X_k being a submatrix of X_ℓ is in accord with this result. Also in accord with this result is the case in which columns of X_k are constructed by linear transformations (e.g., principal components) of a subset of the columns of X_ℓ. When all columns of X_ℓ are also linear functions of all columns of X_k, then $M_\ell \preceq M_k$ and the two models are observationally equivalent. This is the case when, for example, X_k is the matrix of all the principal components of X_ℓ.

In general, the normal linear regression models (5.3) are either nested or partially non-nested, but not globally non-nested (see Pesaran 1987 for more detail). This is not necessarily true when the error variation function for one of the alternative models is not quadratic. For example, consider discriminating between the following models:

$$\begin{aligned} M_k: \quad & y = X_k \beta_k + \varepsilon_k, \qquad \mathcal{V}_k(\varepsilon_k) = |\varepsilon_{ki}| \\ M_\ell: \quad & y = X_\ell \beta_\ell + \varepsilon_\ell, \qquad \mathcal{V}_\ell(\varepsilon_\ell) = \varepsilon^2_{\ell i}. \end{aligned}$$

From Table 1 we find that the ME error distribution for the absolute error variation function is Laplace. The implied ME distribution for y under M_k is Laplace with mean $X_k\beta_k$ and variance ω^2_k. It can be shown (using equation (5.3) of Soofi and Gokhale 1991a) that

$$\bar{K}(\beta_{k0}, \sigma^2_{k0}, \beta^*_\ell, \omega^{2*}_\ell; X_k, X_\ell) = \frac{1}{2} log \left(\frac{\omega^{2*}_\ell}{\sigma^2_{k0}} \right) - \frac{1}{2} log \left(\frac{e}{\pi} \right),$$

where $\omega^{2*}_\ell = \sigma^{2*}_\ell$ is given by (5.4). Therefore, $M_k \preceq M_\ell$ if and only if $\omega^{2*}_\ell / \sigma^2_{k0} = e/\pi$. But this is impossible because $\omega^{2*}_\ell > \sigma^2_{k0}$ by (5.4) and $e < \pi$. Thus, M_k is globally non-nested with respect to M_ℓ. Pesaran (1987) showed a similar result for the case when the distribution of y under M_k is lognormal.

5.1 MDI Tests

Information theoretic testing of linear hypotheses regarding a regression coefficient vector was considered by Kullback and Rosenblatt (1957). They followed the common practice of *assuming* normality of $\boldsymbol{y}$ under various hypotheses and explicated the usual F statistic in terms of the discrimination information function. But the problem can be approached using the ME and MDI developments.

Consider the problem of discriminating between two linear relationships M_1 and M_2 in (5.1) when $X_k = X,\ k = 1, 2,\ \boldsymbol{\beta}_1$ is known, and $\boldsymbol{\beta}_2$ is unknown. Based on quadratic variation function with $\sigma_k^2 = \sigma^2,\ k = 1, 2$, we find two ME models for M_1 and M_2 as shown in (5.3). In this case, the two models are observationally equivalent.

We evaluate (2.8) for the two ME models and find the discrimination information function

$$K(f^*_{\boldsymbol{\beta}_1} : f^*_{\boldsymbol{\beta}_2}; \sigma^2) = \frac{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)' X' X (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}{2\sigma^2}. \tag{5.5}$$

This is the information discrepancy due to the two different means of the two multivariate normal ME distributions under M_1 and M_2.

An information statistic is found by estimating the unknown parameters in (5.5). Kullback and Rosenblatt (1957) suggested replacing the unknown parameters $\boldsymbol{\beta}_2$ and σ^2 in (5.5) with their best unbiased estimates, the ordinary least square estimates. Using $\boldsymbol{b}_2 = \boldsymbol{b}$ in (4.9) and the variance estimate

$$s_2^2 = \frac{(\boldsymbol{y} - X\boldsymbol{b}_2)'(\boldsymbol{y} - X\boldsymbol{b}_2)}{n - p}, \tag{5.6}$$

gives the information statistic

$$\begin{aligned} 2\hat{K}(f^*_{\boldsymbol{\beta}_1} : f^*_{\boldsymbol{\beta}_2}; \sigma^2) &\equiv 2K(f^*_{\boldsymbol{\beta}_1} : f^*_{\boldsymbol{b}_2}; s_2^2) \\ &= \frac{(\boldsymbol{b}_2 - \boldsymbol{\beta}_1)' X' X (\boldsymbol{b}_2 - \boldsymbol{\beta}_1)}{s_2^2} = pF_{p,n-p}. \end{aligned}$$

As indicated in the last expression, the discrimination information statistic follows a multiple of the F distribution with the usual degrees of freedom.

Kullback and Rosenblatt (1957) also developed a discrimination information statistic for discriminating between a normal regression model and its submodel which is a multiple of the usual F ratio. In terms of the ME and MDI developments of this paper, the problem is formulated as follows.

Suppose that $M_1 \preceq M_2$ with columns of X_1 being a subset of columns of X_2. Without loss of generality, let $\boldsymbol{\beta}_1 = (\beta_1, \cdots, \beta_{p_1}, 0, \cdots, 0)'$, $0 \leq p_1 \leq p_2$. Thus $\boldsymbol{\beta}_1$ is partially known and $\boldsymbol{\beta}_2$ is completely unknown.

Based on the quadratic variation with $\sigma_k^2 = \sigma^2$, $k = 1, 2$, the ME model M_k is given by (5.3). According to the result given in Part (ii) of Example 1.2, the normal ME model for M_1 is also the MDI model reference to the normal ME model for M_2, subject to the constraint $E(\boldsymbol{Y}) = X_1\boldsymbol{\beta}_1$.

The MDI statistic for discriminating between the two models is given by

$$\begin{aligned} 2\hat{K}(f^*_{\boldsymbol{\beta}_1} : f^*_{\boldsymbol{\beta}_2}; \sigma^2) &\equiv 2K(f^*_{\boldsymbol{b}_1} : f^*_{\boldsymbol{b}_2}; \hat{\sigma}_2^2) \\ &= \frac{\boldsymbol{b}_2' X_2' X_2 \boldsymbol{b}_2 - \boldsymbol{b}_1' X_1' X_1 \boldsymbol{b}_1}{s_2^2} = (p_2 - p_1)F, \end{aligned} \tag{5.7}$$

where s_2^2 denotes the unbiased variance estimate (5.6) under M_2 as suggested by Kullback and Rosenblatt (1957). The last equation indicates that the MDI statistic (5.7) is a multiple of the usual F ratio obtained by the likelihood ratio method or in the analysis of variance procedure. The MDI derivation of the F ratio is further discussed in the next section.

5.2 MDI Diagnostics

Let $f_k^*(\boldsymbol{y}; \boldsymbol{\theta}_k)$ be the ME distribution implied by the variation function $\mathcal{V}_k(\boldsymbol{\varepsilon}_k)$ associated with the linear relationships M_k in (5.1). Here, $\boldsymbol{\theta}_k$ is $\nu_k \times 1$ vector containing the p_k coefficients and all the parameters related to $\mathcal{V}_k(\boldsymbol{\varepsilon}_k)$. Alternative models are compared according to the minimum information discrepancy between the unknown data-generating distribution $f(\boldsymbol{y})$ and $f_k(\boldsymbol{y}; \boldsymbol{\theta}_k)$.

Among a set of alternatives, M_k, $k = 1, 2, \cdots$, the model M_{k^*} is information optimal if

$$\tilde{K}[f(\boldsymbol{y}) : f^*_{k^*}(\boldsymbol{y}; \tilde{\boldsymbol{\theta}}_{k^*})] \;\leq\; \tilde{K}[f(\boldsymbol{y}) : f^*_k(\boldsymbol{y}; \tilde{\boldsymbol{\theta}}_k)], \quad \textit{for all} \;\; k = 1, 2, \cdots \tag{5.8}$$

where $\tilde{K}$ is an estimate of the MDI function and $\tilde{\boldsymbol{\theta}}_k$ is an estimate of the model parameter. The subscript k of f_k^* underscores the fact that different parametric families may be under consideration.

The partial F ratio is commonly used in search for selecting a submodel M_k of a normal regression model M_L. Examples include selecting the number of lags, stepwise regression, etc. The use of F as a model selection diagnostic sharply differs from the use of F as a test of hypothesis which requires *a priori* specification of a model. The difference between the casual use of F and the formal statistical inference is not generally recognized in common practice.

The MDI statistic (5.7) gives the usual partial F ratio the interpretation of an information criterion for discriminating among the submodels of a normal regression model M_L.

Define $FIC(k)$ as

$$\begin{aligned} FIC(k) &\equiv \frac{2\hat{K}(f^*_{\boldsymbol{\beta}_k} : f^*_{\boldsymbol{\beta}_L}; \sigma^2)}{p_L - p_k} \\ &= \frac{2K(f^*_{\mathbf{b}_k} : f^*_{\mathbf{b}_L}; \hat{\sigma}^2_L)}{p_L - p_k} \\ &= \frac{(n-p_L)(\hat{\sigma}^2_k - \hat{\sigma}^2_L)}{(p_L - p_k)\hat{\sigma}^2_L}, \qquad M_k \preceq M_L. \end{aligned} \tag{5.9}$$

In the last expression, the variance estimates denote the MLE's under the normal models.

According to (5.9), $FIC(k)$ is an estimate of the information loss per variable omitted from the largest model. A submodel M_k is favored over another submodel M_ℓ whenever $FIC(k) < FIC(\ell)$.

The $FIC(k)$ interpretation of the partial F ratio follows from the Kullback and Rosenblatt (1957) derivation of the F ratio in (5.7). This interpretation allows the use of the partial F statistic as a diagnostic for subset selection. Apart from the philosophical issues, in practice, there are generally a number of subsets whose F's are above a threshold; the one with the minimum F will be selected according to the $FIC(k)$ criterion.

In general, estimation of the minimum discrimination information function (5.8) when the data-generating distribution $H[f(y)]$ is unknown is a difficult problem. The equivalent expression (4.1) has been used for estimating the discrimination information function in a number of other contexts and may prove to be useful for the present problem in future research in regression context; see, Soofi et al. (1995) and references therein.

In the regression model selection, the estimation of $K[f(y) : f^*_k(y; \tilde{\boldsymbol{\theta}}_k)]$ is often bypassed and models are compared according to an estimate of the expected log-likelihood function

(4.3). The information criterion (5.8) holds if and only if

$$\tilde{H}_f[f^*_{k^*}(\boldsymbol{y};\tilde{\boldsymbol{\theta}}_{k^*})] \leq \tilde{H}_f[f^*_k(\boldsymbol{y};\tilde{\boldsymbol{\theta}}_k)] \quad \textit{for all} \quad k = 1, 2, \cdots \tag{5.10}$$

Akaike (1973) estimated the expected information discrepancy under the assumption $f(\boldsymbol{y}) = f^*_k(\boldsymbol{y};\boldsymbol{\theta}_L)$ with $\boldsymbol{\theta}'_L = (\boldsymbol{\theta}'_k, \theta_{\nu_k+1}, \cdots, \theta_{\nu_L})$. He developed the following approximate frequentist risk for $\hat{\boldsymbol{\theta}}_k$:

$$2E_{\hat{\boldsymbol{\theta}}_k}\{K[f^*_k(\boldsymbol{y},\boldsymbol{\theta}_L)\} : f^*_k(\boldsymbol{y};\hat{\boldsymbol{\theta}}_k)] \approx -\frac{2}{n} log \frac{f^*_k(\boldsymbol{y};\hat{\boldsymbol{\theta}}_k)}{f^*_k(\boldsymbol{y};\hat{\boldsymbol{\theta}}_L)} + \frac{2\nu_k}{n} - \frac{\nu_L}{n}. \tag{5.11}$$

This approximation is obtained using the second order relationship between relative entropy and Fisher information, the asymptotic normality of the MLE, and asymptotic Chi-square property of some quadratic terms.

Akaike proposed using the approximate risk (5.11) as an estimate of the information criterion (5.8) for selecting a submodel of $f^*_k(\boldsymbol{y};\hat{\boldsymbol{\theta}}_L)$. In a given problem, n, ν_L, and the likelihood function $f^*_k(\boldsymbol{y},\boldsymbol{\theta}_L)$ remain constant. Thus models with various ν_k's are compared according to the information criterion (AIC)

$$AIC(k) = -2log f^*_k(\boldsymbol{y};\hat{\boldsymbol{\theta}}_k) + 2\nu_k, \quad \nu_k = 1, 2, \cdots$$

The quantity, $n^{-1}AIC(k)$ is an almost unbiased estimate of the expected log-likelihood $2\bar{H}_f[f^*_k(\boldsymbol{y};\hat{\boldsymbol{\beta}}_k,\hat{\sigma}^2_k)]$ in (5.10). The model that minimizes $AIC(k)$ is approximately minimum risk and satisfies the MDI criterion (5.8).

For the case of normal regression models (5.3), $\nu_k = p_k + 1$ and $AIC(k)$ is given by

$$\begin{aligned} AIC(k) &= n[log(2\pi e) + log\ \hat{\sigma}^2_k] + 2(p_k + 1) && (5.12) \\ &= 2H[f^*_k(\boldsymbol{y};\hat{\boldsymbol{\beta}}_k,\hat{\sigma}^2_k)] + 2(p_k + 1). && (5.13) \end{aligned}$$

The constants n and $log(2\pi e)$ in expression (5.12) are ignoreable in applications. The expression (5.13) shows that $AIC(k)$ discriminates among alternative normal models by combining the model uncertainty, estimated by the normal regression model entropy, and the number of the parameters k, giving them equal weights.

Sawa (1978) proposed two diagnostics for discriminating among normal regression models based on (5.10). A frequentist diagnostic is obtained by inserting estimates $\hat{\omega}^2$ and $\hat{\sigma}^2$ in the expected log-likelihood component of the approximation (4.11), and is given by

$$BIC(k) = -2log\ f_k^*(y;\hat{\beta}_k, \hat{\sigma}_k^2) + 2(p_k+1)\left(\frac{\hat{\omega}^2}{\hat{\sigma}_k^2}\right) - 2\left(\frac{\hat{\omega}^2}{\hat{\sigma}_k^2}\right)^2.$$

If $\hat{\omega}^2 - \omega^2$ is stochastic of order $n^{-1/2}$, then $BIC(k)$ is an asymptotically unbiased estimate of $2\bar{H}_f[f_k^*(y;\hat{\beta}_k, \hat{\sigma}_k^2)]$ in the risk function (4.11). The issue of inserting an estimate for ω^2 in the risk function has been criticized by. Leamer (1979).

The variance ratio $\hat{\sigma}_k^2/\hat{\omega}^2$ decreases in p_k and is interpreted by Sawa as a "discounting factor" for the penalty of increasing the number of variables. For $\hat{\sigma}_k^2/\hat{\omega}^2 = 1$, $BIC(k)$ reduces to $AIC(k)$. Estimation of ω^2 is the main problem with the implementation of $BIC(k)$

For the normal regression models $BIC(k)$ is

$$\begin{aligned} BIC(k) &= n[log(2\pi e) + log\ \hat{\sigma}_k^2] + 2(p_k+1)\left(\frac{\hat{\omega}^2}{\hat{\sigma}_k^2}\right) - 2\left(\frac{\hat{\omega}^2}{\hat{\sigma}_k^2}\right)^2 \\ &= 2H[f_k^*(y;\hat{\beta}_k, \hat{\sigma}_k^2)] + 2(p_k+1)\left(\frac{\hat{\omega}^2}{\hat{\sigma}_k^2}\right) - 2\left(\frac{\hat{\omega}^2}{\hat{\sigma}_k^2}\right)^2. \end{aligned}$$

For discriminating between two nested normal models $M_k \preceq M_L$ with columns of X_k being a subset of columns of X_L, Sawa suggested using $\hat{\omega}^2 = \hat{\sigma}_L^2$. Under the assumption of $f(y) = N(\mu, \omega^2 I_n)$, model selection based on $BIC(k)$ is equivalent to that based on the magnitude of $FIC(k)$ statistic defined in (5.9). The submodel M_k is favored over M_L whenever $BIC(k) < BIC(L)$, with the condition in terms of the variance ratio is given as

$$nlog\left(\frac{\hat{\sigma}_L^2}{\hat{\sigma}_k^2}\right) - 2(p_k+2)\left(\frac{\hat{\sigma}_L^2}{\hat{\sigma}_k^2}\right) + 2\left(\frac{\hat{\sigma}_L^2}{\hat{\sigma}_k^2}\right)^2 + 2(p_L+1) < 0.$$

The $BIC(k)$ decision rule is based on the magnitude of $FIC(k)$ due the fact that

$$\frac{\hat{\sigma}_k^2}{\hat{\sigma}_L^2} = 1 + \frac{p_L - p_k}{n - p_L} FIC(k).$$

Sawa (1978) also developed a Bayesian diagnostic for discriminating between nested normal regression models based on the lower bound (4.10). He found that the Bayes estimate that, under M_k, minimizes the lower bound in (4.11) is

$$\tilde{\sigma^*}_k^2 = \frac{n + p_k}{n - p_k - 2}\hat{\sigma}_k^2.$$

The Bayes estimate gives the minimum attainable risk

$$2E_{(\boldsymbol{\beta}_k,\sigma_k^2)|\boldsymbol{y}}\left\{\bar{H}_f[f_k^*(\boldsymbol{y};\tilde{\boldsymbol{\beta}}_k^*,\tilde{\sigma^*}_k^2)]\right\} = -\frac{2}{n}log\ f_k^*(\boldsymbol{y};\hat{\boldsymbol{\beta}}_k,\hat{\sigma}_k^2) + log\left(\frac{n+p_k}{n-p_k-2}\right). \tag{5.14}$$

For the case of two nested normal models $M_k \preceq M_L$ with columns of X_k being a subset of columns of X_L, Sawa showed that the reduced model is favored by the minimum attainable risk (5.14) if and only if

$$FIC(k) \leq \frac{2(n-1)(n-p_L)}{(n+p_k)(n-p_L-2)}.$$

Young (1987) developed an information criterion for discriminating between normal regression models by defining risk as $E_{(\boldsymbol{\mu},\omega^2)|\boldsymbol{y}}E_{(\tilde{\boldsymbol{\beta}},\sigma^2)|\boldsymbol{y}}\ K[f(\boldsymbol{y};\boldsymbol{\mu},\omega) : f(\boldsymbol{y};\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)]$. Young assumed $f(\boldsymbol{y}) = N(\boldsymbol{\mu},\omega^2\boldsymbol{I}_n)$, and used the following prior distributions for the parameters:

$$\begin{aligned} \pi(\boldsymbol{\mu}|\omega^2) &= N(\boldsymbol{m},\omega^{-2}W^{-1}), & \pi(\omega^{-2}) = Gamma(\alpha,\nu) \\ \pi(\boldsymbol{\beta}_k|\sigma_k^2) &= N(\boldsymbol{m}_k,\sigma_k^{-2}W_k^{-1}), & \pi(\sigma_k^{-2}) = Gamma(\alpha_k,\nu_k). \end{aligned} \tag{5.15}$$

When the priors are weak (i.e., $W \to 0,\ \ \alpha \to 0,\ \ \nu \to 0,\ \ W_k \to 0,\ \ \alpha_k \to 0,\ \ \nu_k \to 0$), the Bayes estimates of $\boldsymbol{\beta}_k$ and σ_k^2 are approximately equal to the MLE under the model M_k. If the priors are weak, then the risk is approximately minimized by the model that minimizes

$$CIC(k) = n\ log\hat{\sigma}_k^2 + p_k.$$

Comparing with expression (5.12), we note that $CIC(k)$ gives one half as much weight to the dimension of the model as that given by $AIC(k)$.

5.3 Other Discrimination Information Diagnostics

Ibrahim and Laud (1994) used discrimination information function for model selection in the context of a Bayesian predictive approach. In this approach, the model M_k is evaluated based on the predictive density for a set of n new observations. Let $\boldsymbol{y}_N[X_k]$ denote the vector of new observations taken at the design matrix X_k. Then using the normal ME distribution (5.3) as the likelihood function under M_k, the predictive density is given by

$$f(\boldsymbol{y}_N[X_k]|\boldsymbol{y}) = \int\int f(\boldsymbol{y}|\boldsymbol{\beta}_k,\sigma_k^2)\pi(\boldsymbol{\beta}_k,\sigma_k^2|\boldsymbol{y})d\boldsymbol{\beta}_k\ d\sigma_k^2.$$

These authors considered the normal-gamma prior (5.15) for the model parameters. The prior mean for $\boldsymbol{\beta}_k$ was chosen by the pseudo-parameter value $\boldsymbol{m}_k = (X_k'X_k)^{-1}X_k'\boldsymbol{\mu}_0$ where $\boldsymbol{\mu}_0$ is a "guess" value for $\boldsymbol{\mu}$. The prior precision matrix was chosen as $W_k = X_k'X_k$.

Alternative normal models are compared with the largest model M_L according to the symmetrized discrimination information function

$$K_{k,L} = K[f(\boldsymbol{y}_N[X_k]|\boldsymbol{y}) : f(\boldsymbol{y}_N[X_L]|\boldsymbol{y})] + K[f(\boldsymbol{y}_N[X_L]|\boldsymbol{y}) : f(\boldsymbol{y}_N[X_k]|\boldsymbol{y})].$$

Computation of this expression involves the evaluation of the discrimination function between two multivariate t distributions which does not have a closed form. Ibrahim and Laud found an approximate expression for the symmetrized information function and showed that for the case of vague priors, it is a monotone function of $FIC(k)$.

Carota, Parmigiani, and Polson (1996) used discrimination information function in the context of "model elaboration". In their approach, a model M is embedded in a larger family of models M_ζ; i.e., $M = M_{\zeta_0}$ for a specific value ζ_0. The discrimination information between the posterior and prior of the elaboration parameter $K[\pi(\zeta|\boldsymbol{y}) : \pi(\zeta)]$, is used as the diagnostic for the elaboration. When $K[\pi(\zeta|\boldsymbol{y}) : \pi(\zeta)]$ is small, the elaborated model is not supported by the data. These authors developed the following linearized approximation

$$K_L[\pi(\zeta|\boldsymbol{y}) : \pi(\zeta)] \approx log(B) + S\ E_{\zeta|\boldsymbol{y}}(\zeta - \zeta_0),$$

where B is the Savage density ratio which is equivalent to the Bayes factor under certain conditions and S is the score function defined as follows:

$$B = \frac{f(\boldsymbol{y}|M_{\zeta_0})}{f(\boldsymbol{y})}, \qquad S = \frac{\partial}{\partial\zeta} log f(\boldsymbol{y}|\zeta)\bigg|_{\zeta=\zeta_0}.$$

These authors discussed a regression example in which the elaboration is defined by the inclusion of an additional variable in the model. That is, the elaboration parameter is the coefficient of the additional variable. The normal likelihood function (5.3) and the normal-gamma prior (5.15) are used. In the prior, the elaboration parameter has mean zero and is uncorrelated with the other coefficients. In this problem, $K[\pi(\zeta|\boldsymbol{y}) : \pi(\zeta)]$ is also the discrimination information between two Student-t distributions and does not have a closed form. Carota, Parmigiani, and Polson (1996) showed that the linearized version provides an accurate approximation when compared with the case of known error variance.

5.4 Complexity Diagnostics

We have already seen that the MDI diagnostics $AIC(k)$, $BIC(k)$, and $CIC(k)$ discriminate among the alternative models based on model fit as indicated by the log-likelihood term and the model complexity as indicated by a term involving p_k; see also Poskitt (1987). Some authors (Rissanen 1986, 1987a, 1987b; Bozdogan 1990; Bozdogan and Haughton 1995) have proposed model selection criteria with emphasis of model complexity.

Rissanen defined stochastic complexity for a given class of models as "the number of binary digits with which the observations can be described" (Rissanen 1987a). Let the model class be defined by the pair of density functions:

$$\mathcal{C}_{f,\pi} = \{[f(y|p_k, \boldsymbol{\theta}_k),\ \pi(\boldsymbol{\theta}_k|p_k)],\ \ p_k = 0, 1, \cdots\},$$

where p_k is the number of free parameters in the model pair. Then *stochastic complexity* of the data points $y_1, \cdots, y_n$ is measured by

$$SC(k) = -log \sum_{k=0}^{p_k} \frac{1}{p_k + 1} \int f(y|p_k, \boldsymbol{\theta}_k) d\pi(\boldsymbol{\theta}_k|p_k), \quad p_k < n. \tag{5.16}$$

The prior is assumed to be concentrated near the MLE estimate $\hat{\boldsymbol{\theta}}_k$. The model that minimizes (5.16) in $\mathcal{C}_{f,\pi}$ is preferred.

For sufficiently large n, an approximate upper bound for $SC(k)$ is minimized and the criterion is referred to as the *Minimum Description Length (MDL)*. Ignoring the non-essential terms, $MDL(k)$ compares the models according to:

$$\begin{aligned} MDL(k) &\approx -logf(y|\hat{\boldsymbol{\theta}}_k) + \frac{1}{2} log \left| \frac{\partial^2 log f(y|\hat{\boldsymbol{\theta}}_k)}{\partial \hat{\boldsymbol{\theta}}_k^2} \right| \\ &\approx -logf(y|\hat{\boldsymbol{\theta}}_k) + \frac{p_k}{2} log\ n. \end{aligned}$$

For the normal regression model, $MDL(k)$ may be written as

$$MDL(k) \approx n[log(2\pi e) + log\ \hat{\sigma}_k^2] + \frac{log\ n}{2}\ p_k.$$

Thus, the weight $(log\ n)/2$ given to the dimension of the model p_k is larger for $MDL(k)$ in comparison with the MDI diagnostics.

Bozdogan (1990) defined the complexity of a p-variate normal distribution with covariance Σ by the maximal mutual information

$$\begin{aligned} C[\Sigma(\boldsymbol{X})] &\equiv \max_{T\in\mathcal{T}} \vartheta[T\boldsymbol{X} \wedge (T\boldsymbol{X})_1, \cdots, (T\boldsymbol{X})_p] \\ &= \frac{p}{2} log \left(\frac{Tr(\Sigma)}{p}\right) - \frac{1}{2} log\ |\Sigma|, \end{aligned} \tag{5.17}$$

where $\mathcal{T}$ is the set of all orthonormal transformations in $\Re^p$. This measure is motivated by the fact that the mutual information $\vartheta(\boldsymbol{X} \wedge X_1, \cdots, X_p)$ is not invariant under axis rotations (see Example 2.3), whereas (5.17) is invariant under axis rotations.

Bozdogan and Haughton (1995) defined *Informational Complexity* of a model as

$$ICOMP(k) = -2 log f(\boldsymbol{y}; \boldsymbol{\theta}_k) + 2C[\Sigma(\hat{\boldsymbol{\theta}}_k)],$$

where $\Sigma(\hat{\boldsymbol{\theta}}_k)$ is the covariance matrix of the estimated parameters. These authors also discussed attaching a weight a_n to the complexity term.

Two alternative methods proposed by Bozdogan and Haughton (1995) for estimating the complexity term. One method uses a sample version of the covariance matrix, $\hat{\Sigma}(\hat{\boldsymbol{\theta}}_k)$. A second method uses the inverse of the estimated Fisher information matrix $[\hat{\mathcal{F}}(\hat{\boldsymbol{\theta}})]^{-1}$. In this case, the informational complexity criterion is

$$ICOMPIFIM(k) = -2 log f(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_k) + 2C[\{\hat{\mathcal{F}}(\hat{\boldsymbol{\theta}}_k)\}^{-1}].$$

For the normal regression model, $ICOMP(k)$ is computed as

$$\begin{aligned} ICOMP(k) &= -2 log[f_k^*(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k^2)] + 2C[\hat{\Sigma}(\hat{\boldsymbol{\beta}}_k)] \\ &= n[log(2\pi e) + log\ \hat{\sigma}_k^2] + p_k log \left(\frac{Tr(X_k' X_k)^{-1}}{p_k + 1}\right) - log\ |X_k' X_k|^{-1}. \end{aligned}$$

The second version of the complexity criterion is computed as

$$\begin{aligned} ICOMPIFIM(k) &= n[log(2\pi e) + log\ \hat{\sigma}_k^2] \\ &\quad + (p_k + 1) log \left(\frac{Tr(X_k' X_k)^{-1} + 2\hat{\sigma}_k^2/n}{p_k + 1}\right) \\ &\quad - log\ |X_k' X_k|^{-1} - log \left(\frac{2\hat{\sigma}_k^2}{n}\right). \end{aligned}$$

6 Collinearity Analysis

The existence of a near linear relationship among the columns of the regression matrix X is referred to as (near) collinearity. When X is collinear, the inversion of $X'X$ is problematic and creates computational and conceptual problems in regression analysis. The computational issue is that the solution to the least square equations, $\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y}$, changes drastically with a slight perturbation of X. The conceptual aspects of the collinearity are the problems associated with the inference based on a distribution that depends on a collinear regression matrix.

Often the subject of inference is the regression coefficient vector, $\boldsymbol{\beta}$. The traditional literature has casted the collinearity problem as the lack of adequate "information" in the data for estimating $\boldsymbol{\beta}$, but has not gone beyond the semantic notion of information. Formally, the effects of collinearity on *information* about $\boldsymbol{\beta}$ can be measured in terms of the entropy, relative entropy, and mutual information functions discussed in Section 2 (Soofi 1988, 1990).

Consider the normal ME regression model $f^*(\boldsymbol{y};\boldsymbol{\beta},\sigma^2) = N(X\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n)$. In the Bayesian framework, $\boldsymbol{\beta}$ is subject to variation and the posterior distribution of $\boldsymbol{\beta}$, given the data is the vehicle of inference about the regression coefficients. In the frequentist approach, the inference about the regression coefficient is based on the distribution induced by the sampling variation of the data on an estimate of $\boldsymbol{\beta}$. However, as will be seen, there are some information dualities between the two approaches.

The following reparametrization of (3.1) is useful for collinearity analysis.

$$\boldsymbol{y} = (X\Gamma)(\Gamma'\boldsymbol{\beta}) + \boldsymbol{\varepsilon} = W\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{6.1}$$

where $\Gamma = [\gamma_1,\cdots,\gamma_p]$ is the orthogonal matrix of the eigenvectors of $X'X$ and $W = [W_1,\cdots,W_p]$ is the transformed regression matrix in the directions of the principal components of X. Note that

$$W'W = \Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}, \tag{6.2}$$

$\lambda_1 \geq, \cdots, \geq \lambda_p$ being the eigenvalues of $X'X$.

6.1 Collinearity Diagnostics for Least Square Regression

The solution to the least square equation, $\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y}$, is justified as an estimate of $\boldsymbol{\beta}$ according to various estimation criteria. It is the MLE of $\boldsymbol{\beta}$ under normal ME regression model. In Section 4, $\boldsymbol{b}$ was seen as the MDIM, an approximate MDIB, and the BMOM estimate of $\boldsymbol{\beta}$.

Consider the case when the only prior information used about $\boldsymbol{\beta}$ is the ranges of the variations of β_j's. As shown in Table 2, the ME prior $\pi^*(\boldsymbol{\beta})$ is uniform. The prior entropy is an increasing function of the ranges of the variations. If the ranges are very large, then the prior is noninformative about predicting the regression coefficients; i.e., $H[\pi^*(\boldsymbol{\beta})]$ is a large constant. Given the data and σ^2, the conditional posterior distribution is $\pi(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2) = N[\boldsymbol{b}, \sigma^2(X'X)^{-1}]$.

The amount of uncertainty in predicting a value of $\boldsymbol{\beta}$ is measured by the posterior entropy

$$H_X(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2) = \frac{p}{2} log(2\pi e\sigma^2) - log|X'X|^{1/2}. \tag{6.3}$$

Since the prior is noninformative, the effects of collinearity on the information content of the data about the regression coefficients are examined based on $I_X(\boldsymbol{\beta}|\sigma^2) = -H_X(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2)$.

The advantage of the representation (6.1) is that $\pi(\boldsymbol{\alpha}|\boldsymbol{y}, \sigma^2) = N(\boldsymbol{a}, \sigma^2\Lambda^{-1})$, $\boldsymbol{a} = \Gamma'\boldsymbol{b}$ and Λ is defined in (6.2). That is,the regression coefficients in the directions of the principal components $\alpha_j = \gamma_j'\boldsymbol{\beta}$ are uncorrelated normal, hence are independent. Writing the determinant in (6.3) as the product of the eigenvalues, the posterior entropy is decomposed in terms of the entropies of the independent components of $\boldsymbol{\alpha}$,

$$\begin{aligned} I_X(\boldsymbol{\beta}|\sigma^2) &= \sum_{j=1}^{p} log\ \lambda_j^{1/2} - \frac{p}{2} log(2\pi e\sigma^2) \\ &= \sum_{j=1}^{p} I_X(\gamma_j'\boldsymbol{\beta}|\sigma^2) \\ &= \sum_{j=1}^{p} I_{W_j}(\alpha_j|\sigma^2) \\ &= I_W(\boldsymbol{\alpha}|\sigma^2). \end{aligned} \tag{6.4}$$

The components of information in (6.4) are comparable only when the columns of X are equilibrated. Henceforth, assume that the columns of X are scaled so that $X'X$ is in correlation form.

Given the error variance σ^2, the information quantities $I_{W_j}(\alpha_j|\sigma^2)$ are ordered according to the eigenvalues of $X'X$, with the first element $\alpha_1 = \boldsymbol{\gamma}_1'\boldsymbol{\beta}$ being the most informative (minimum entropy); i.e. the least difficult to predict. Note that "given σ^2" implies the presence of all the components in the regression equation. Thus the components of information in (6.4) are not criteria for reducing the model. They display the information spectrum of the regression matrix as a whole.

The explanatory variables are most informative (minimum entropy) about the regression coefficients when the regression matrix is orthogonal. Thus, a measure of *information loss due to collinearity* of X is given by the information difference

$$\begin{aligned} IL_X(\boldsymbol{\beta}|\sigma^2) &= I_{max}(\boldsymbol{\beta}|\sigma^2) - I_X(\boldsymbol{\beta}|\sigma^2) \\ &= I_{X^o}(\boldsymbol{\beta}|\sigma^2) - I_X(\boldsymbol{\beta}|\sigma^2) \\ &= -log|X'X|^{1/2} \\ &= -\sum_{j=1}^{p} log\ \lambda_j^{1/2}, \end{aligned} \tag{6.5}$$

where X^o denote an orthogonal reference regression matrix.

The *information indices* of a regression matrix are defined by the information differences:

$$\begin{aligned} \Delta_j(X) &= I_X(\boldsymbol{\gamma}_1'\boldsymbol{\beta}|\sigma^2) - I_X(\boldsymbol{\gamma}_j'\boldsymbol{\beta}|\sigma^2) \\ &= log\left(\frac{\lambda_1}{\lambda_j}\right)^{1/2} \\ &= log\ \kappa_j(X), \quad j = 1, \cdots, p, \end{aligned}$$

where $\kappa_j(X),\ j = 1, \cdots, p$ are the condition indices of the regression matrix (Belsley, Kuh, and Welsch 1980).

The *information number* of a regression matrix is defined by the information range

$$\Delta(X) = log\left(\frac{\lambda_1}{\lambda_p}\right)^{1/2} = log\ \kappa(X),$$

where $\kappa(X)$ is the condition number of X.

The information spectrum of X^o is uniform and $\Delta_j(X^o) = 0$ for all $j = 1, \cdots, p$. Therefore, the origin of measurement for the information indices is the orthogonality. For example, $\Delta(X)$ directly measures the maximum extent of the collinearity of X in relationship to the orthogonality. Other consequences of the logarithmic transformations of $|X'X|$ and κ are discussed Soofi (1990).

In the least square regression, a computed $\boldsymbol{b}$ is viewed as an outcome of the sampling distribution $f(\boldsymbol{b}|\boldsymbol{\beta},\sigma^2) = N[\boldsymbol{\beta},\sigma^2(X'X)^{-1}]$. The entropy of the sampling distribution is also given by (6.3). Therefore, all the statements made regarding the posterior entropy (6.3) also have frequentist interpretations. In the sampling theory inference, (6.3) quantifies the amount of uncertainty in predicting a value of the least square estimate $\boldsymbol{b}$.

The effects of collinearity on the least square estimation may also be measured by the discrimination information function between the actual sampling distribution $f_X(\boldsymbol{b}|\boldsymbol{\beta},\sigma^2) = N[\boldsymbol{\beta},\sigma^2(X'X)^{-1}]$ and the sampling distribution of the estimate as if the regressors were orthogonal. In collinearity analysis, traditionally it is assumed that the artificial reference sampling distribution based on the orthogonal regression matrix has mean $\boldsymbol{\beta}$, i.e., $f_{X^o}(\boldsymbol{b}^o|\boldsymbol{\beta},\sigma^2) = N(\boldsymbol{\beta},\sigma^2\boldsymbol{I}_p)$. Consequently, the discrimination information between the two sampling distributions is given by the information discrepancy due to the covariances of two normal distributions:

$$\begin{aligned} K(\boldsymbol{b}:\boldsymbol{b}^o) &\equiv K[f_X(\boldsymbol{b}|\boldsymbol{\beta},\sigma^2) : f_{X^o}(\boldsymbol{b}^o|\boldsymbol{\beta},\sigma^2)] \\ &= \frac{1}{2}\left[Tr(X'X)^{-1} - log|X'X|^{-1} - p\right] \\ &= \frac{1}{2}\left[\sum_{j=1}^{p}\frac{1}{\lambda_j} + \sum_{j=1}^{p} log\ \lambda_j - p\right]. \end{aligned}$$

Since $K(\boldsymbol{b}:\boldsymbol{b}^o) \geq 0$, with equality if and only if X is orthogonal and $K(\boldsymbol{b}:\boldsymbol{b}^o) \longrightarrow \infty$ as X descents to perfect collinearity, $K(\boldsymbol{b}:\boldsymbol{b}^o)$ measures the *loss of information* in the least square estimation due to the nonorthogonality of X.

$Tr(X'X)^{-1}$ and $|X'X|$ are collinearity diagnostics with traditional statistical interpretations. $K(\boldsymbol{b}:\boldsymbol{b}^o)$ is composed of the trace, determinant, and the rank of $(X'X)^{-1}$. The information loss is measured by a comprehensive summary of the covariance matrix of the sampling distribution and is inclusive of the traditional measures .

The discrimination information function between the actual posterior distribution of the regression coefficient vector, $\pi_X(\boldsymbol{\beta}|\boldsymbol{y},\sigma^2) = N[\boldsymbol{b},\sigma^2(X'X)^{-1}]$, and the posterior distribution as if the regressors were orthogonal, $\pi_{X^o}(\boldsymbol{\beta}|\boldsymbol{y},\sigma^2) = N(\boldsymbol{b}^o, \sigma^2\boldsymbol{I}_p)$ is

$$K(\pi_X : \pi_{X^o}|\boldsymbol{y},\sigma^2) = K(\boldsymbol{b} : \boldsymbol{b}^o) + \frac{(\boldsymbol{b}-\boldsymbol{b}^o)'(\boldsymbol{b}-\boldsymbol{b}^o)}{2\sigma^2}.$$

The second term is the information discrepancy due to two different posterior means. This term is a measure of the effect of collinearity on the solutions to the least square equations which is ignored in the traditional collinearity diagnostics.

It is well known that when X is near-collinear, the least square solutions are sensitive to small perturbations of X. The discrimination information function between the perturbed posterior distribution and the actual posterior distribution is

$$\begin{aligned} K(\pi_{X^*} : \pi_X|\boldsymbol{y},\sigma^2) &= \frac{1}{2}\left[Tr(X'X)(X^{*\prime}X^*)^{-1} - log|X'X(X^{*\prime}X^*)^{-1}| - p\right] \\ &\quad + \frac{(\boldsymbol{b}-\boldsymbol{b}^*)'X'X(\boldsymbol{b}-\boldsymbol{b}^*)}{2\sigma^2}, \end{aligned}$$

where $X^* = X + dX$ is the perturbed regression matrix and $\boldsymbol{b}^*$ is the perturbed least square solution. The first term measures the effect of the perturbation on the covariance structure. This term may also be interpreted as the effect of the perturbation on the sampling distribution of the least square estimate. The second term is the effect of perturbation on the solutions to the least square equations.

6.2 Collinearity Diagnostics With Prior Information

Consider the case when the prior information assumed about the regression coefficients is in the form of the quadratic variation functions $\mathcal{V}(\boldsymbol{\beta}) = (\beta_j - m_j)$, $j = 1, \cdots, p$. Table 2 gives the ME prior $\pi^*(\boldsymbol{\beta}|\boldsymbol{m},\tau^2) = N(\boldsymbol{m}, \tau^2\boldsymbol{I}_p)$. The prior independence among the regression coefficients is due to the fact that no information about the interrelationships between were used in the ME computation.

The prior uncertainty about the regression coefficients is given by

$$H(\boldsymbol{\beta}|\tau^2) = \frac{p}{2}log(2\pi e) + \frac{p}{2}log\ \tau^2. \tag{6.6}$$

Based on the ME normal likelihood, the posterior distribution given the error variance is $\pi(\boldsymbol{\beta}|\boldsymbol{y},\sigma^2,\boldsymbol{m},\tau^2) = N[\boldsymbol{b}(\phi,\boldsymbol{m}), \sigma^2(X'X+\phi\boldsymbol{I}_p)^{-1}]$ where

$$b(\phi, \boldsymbol{m}) = (X'X + \phi \boldsymbol{I}_p)^{-1}(X'\boldsymbol{y} + \phi \boldsymbol{m}). \tag{6.7}$$

and

$$\phi = \frac{\sigma^2}{\tau^2}$$

is the prior to model precision ratio.

The posterior entropy is

$$H_X(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \tau^2) = \frac{p}{2} log(2\pi e) + \frac{p}{2} log\ \sigma^2 + log|\phi \boldsymbol{I}_p + X'X|^{1/2}. \tag{6.8}$$

The sample information about the regression coefficients is given by the entropy difference

$$\begin{aligned} \vartheta_X(\boldsymbol{\beta}|\phi) &= H(\boldsymbol{\beta}|\tau^2) - H_X(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \tau^2) \\ &= log|\boldsymbol{I}_p + \phi^{-1}X'X|^{1/2}. \end{aligned} \tag{6.9}$$

In fact, $\vartheta_X(\boldsymbol{\beta}|\phi)$ is the mutual information between $\boldsymbol{y}$ and $\boldsymbol{\beta}$, $\vartheta_X(\boldsymbol{\beta}|\phi) = \vartheta_X(\boldsymbol{\beta} \wedge \boldsymbol{y}|\phi)$. Here, taking the expectation with respect to the distribution of $\boldsymbol{y}$ is not needed because the entropy difference (6.9) is functionally independent of $\boldsymbol{y}$. (More generally, *every* sample drawn from a normal distribution is informative about the mean which is also drawn from a normal distribution.)

Although the prior entropy (6.6) depends on the prior variance and the posterior entropy (6.8) depends on the prior and error variances, the sample information (6.9) depends on the precision ratio ϕ, which is the pivotal quantity in the collinearity analysis in the presence of prior information. The sample information is decomposable as:

$$\begin{aligned} \vartheta_X(\boldsymbol{\beta}|\phi) &= \sum_{j=1}^{p} log(1 + \phi^{-1}\lambda_j)^{1/2} \\ &= \sum_{j=1}^{p} \vartheta_{W_j}(\boldsymbol{\alpha}_j|\phi) \\ &= log|\boldsymbol{I}_p + \phi^{-1}\Lambda|^{1/2} \\ &= \vartheta_W(\boldsymbol{\alpha}|\phi). \end{aligned} \tag{6.10}$$

Thus given ϕ, the components of the sample information $\vartheta_{W_j}(\boldsymbol{\alpha}_j|\phi)$ are ordered according to the eigenvalues.

The mutual information is maximum when the regression matrix is orthogonal. In the presence of prior information, a measure of *information loss due to collinearity* of X is given by the information difference

$$\begin{aligned} \vartheta L_X(\boldsymbol{\beta}|\phi) &= \vartheta_{max}(\boldsymbol{\beta}|\phi) - \vartheta_X(\boldsymbol{\beta}|\phi) \\ &= \vartheta_{X^\circ}(\boldsymbol{\beta}|\phi) - \vartheta_X(\boldsymbol{\beta}|\phi) \\ &= \sum_{j=1}^{p} log \left(\frac{\phi+1}{\phi+\lambda_j}\right)^{1/2} \end{aligned}$$

The sample information loss $\vartheta L_X(\boldsymbol{\beta}|\phi)$ has the following properties.

(i) $\vartheta L_X(\boldsymbol{\beta}|\phi)$ is monotonically decreasing in ϕ. Given σ^2, $\vartheta L_X(\boldsymbol{\beta}|\phi)$ is monotonically decreasing in the prior precision τ^{-2}.

(ii) Given σ^2, $\vartheta L_X(\boldsymbol{\beta}|\phi) < IL_X(\boldsymbol{\beta}|\sigma^2)$ for all $\tau^2 > 0$; and $\vartheta L_X(\boldsymbol{\beta}|\phi) \to IL_X(\boldsymbol{\beta}|\sigma^2)$ as $\tau^2 \to \infty$.

Given the precision ratio ϕ, the information indices of X are obtained by the information differences

$$\begin{aligned} \Delta_j(X,\phi) &= \vartheta_X(\gamma_1'\boldsymbol{\beta}|\phi) - \vartheta_X(\gamma_j'\boldsymbol{\beta}|\phi) \\ &= H_X(\gamma_j'\boldsymbol{\beta}|\phi) - H_X(\gamma_1'\boldsymbol{\beta}|\phi) \\ &= log\left(\frac{\phi+\lambda_1}{\phi+\lambda_j}\right)^{1/2} \\ &= log\ \kappa_j[\phi \boldsymbol{I}_p + X'X], \quad j=1,\cdots,p, \end{aligned} \tag{6.11}$$

$\Delta_j(X,\phi)$ generalizes $\Delta_j(X)$ which is given by $\phi = 0$. Further generalization may be obtained by using ϕ_j, $j = 1,\cdots,p$ in (6.11).

The information indices (6.11) are also interpretable in the sampling theory framework. By letting $\boldsymbol{m} = \mathbf{0}$ in (6.7) we obtain the ridge estimate of the regression coefficients with the ridge parameter, ϕ. The entropy of the sampling distribution of the ridge estimate is also given by (6.8). By the second equality in (6.11) the information indices display the information spectrum of the sampling distribution of the ridge estimate $\boldsymbol{b}(\phi)$.

Another measure of information loss due collinearity in the presence of prior information is given by the discrimination information function between the actual posterior distribution $\pi_X(\boldsymbol{\beta}|\boldsymbol{y},\sigma^2,\boldsymbol{m},\tau^2)$ and the posterior distribution as if the regressors were orthogonal,

$\pi_{X^o}(\boldsymbol{\beta}|\boldsymbol{y},\sigma^2,\boldsymbol{m},\tau^2)$. This discrimination information loss due to collinearity is given by

$$K(\pi_X : \pi_{X^o}|\boldsymbol{y},\sigma^2,\boldsymbol{m},\tau^2) = \\ \frac{1}{2}\left\{Tr\left[(1+\phi)(X'X+\phi\boldsymbol{I}_p)^{-1}\right] - log\left|(1+\phi)(X'X+\phi\boldsymbol{I}_p)^{-1}\right| - p\right\} \\ +\frac{(1+\phi)}{2\sigma^2}[\boldsymbol{b}(\phi,\boldsymbol{m})-\boldsymbol{b}^o(\phi,\boldsymbol{m})]'[\boldsymbol{b}(\phi,\boldsymbol{m})-\boldsymbol{b}^o(\phi,\boldsymbol{m})].$$

The discrimination information loss has the following properties.

(i) For all $\phi > 0$, $K(\pi_X : \pi_{X^o}|\boldsymbol{y},\sigma^2,\boldsymbol{m},\tau^2)$ is finite for all X.

(ii) For a given X, $K(\pi_X : \pi_{X^o}|\boldsymbol{y},\sigma^2,\boldsymbol{m},\tau^2)$ is monotonically decreasing in ϕ.

(iii) For all $\tau^2 \geq 0$, $K(\pi_X : \pi_{X^o}|\boldsymbol{y},\sigma^2,\boldsymbol{m},\tau^2) \leq K(\pi_X : \pi_{X^o}|\boldsymbol{y},\sigma^2)$.

We have seen that the information loss due to collinearity, measured by $\vartheta L_X(\boldsymbol{\beta}|\phi)$ or $K(\pi_X : \pi_{X^o}|\boldsymbol{y},\sigma^2,\boldsymbol{m},\tau^2)$ is always less than the loss when no prior information regarding the variation of the regression coefficients is used. Therefore, one can compensate for the sample loss of information due to collinearity by acquiring nonsample information in order to decrease the maximum average variation of the regression coefficients, τ^2; i.e. to increase the prior precision. A collinearity information graph may be constructed by plotting various information functions against ϕ or τ^2, see Soofi (1990). The graph is useful for determining the prior precision needed for a certain amount of collinearity loss reduction,

Note that the form of prior information used is important for the collinearity analysis. The normal ME prior $\pi^*(\boldsymbol{\beta}|\boldsymbol{m},\tau^2) = N(\boldsymbol{m},\tau^2\boldsymbol{I}_n)$ is obtained based on the information regarding the variations of the individual regression coefficients. The diagnostics discussed above are based on the prior ignorance about the interrelationships among the regression coefficients. If we wish to include information regarding the interrelationships among the regression coefficients in the prior, we should use the variation function of the form $\mathcal{V}(\boldsymbol{\beta}) = (\boldsymbol{\beta}-\boldsymbol{m})(\boldsymbol{\beta}-\boldsymbol{m})'$. Then Table 2 gives the ME prior $\pi^*(\boldsymbol{\beta}|\boldsymbol{m},\tau^2,\Psi) = N(\boldsymbol{m},\tau^2\Psi)$. In this case, the sample information about the regression coefficients is given by

$$\vartheta_X(\boldsymbol{\beta}|\phi,\Psi) = log|\boldsymbol{I}_p + \phi^{-1}\Psi X'X|^{1/2}.$$

As an example, consider Zellner's g-prior (3.6) which uses $\Psi = (X'X)^{-1}$. Based on this prior, the sample information about the regression coefficients is given by

$$\vartheta_X(\beta|\phi,\Psi) = (X'X)^{-1}) = \frac{p}{2} log\left(\frac{1+\phi}{\phi}\right).$$

Because the sample information regarding the correlation structure of the regression coefficients has already been used in the prior, the sample does not add any new information about the correlation. This fact is reflected in the sample information function being free from X. The two cases, $\Psi = \boldsymbol{I}_p$, and $\Psi = (X'X)^{-1}$ are two extremes in their impacts on the collinearity effects. The prior independence reduces the effects collinearity on the information about the regression coefficients, whereas $\Psi = (X'X)^{-1}$ leaves the collinearity problem intact.

6.3 A Collinearity Diagnostic for Random Regressors

Consider the case when the explanatory variables $\boldsymbol{X} = (X_1, \cdots, X_p)'$ are jointly normal. Then the information about each X_j provided by the set of other explanatory variables $\boldsymbol{X}_{(-j)}$ is given by the entropy reduction

$$\begin{aligned} \vartheta(X_j|\boldsymbol{X}_{(-j)}) &= H(X_j) - H(X_j|\boldsymbol{X}_{(-j)}) \\ &= -\frac{1}{2} log[1-\rho^2(X_j;\boldsymbol{X}_{(-j)})], \quad j = 1, \cdots, p, \end{aligned}$$

where $\rho^2(X_j;\boldsymbol{X}_{(-j)})$ is the square of multiple correlation between X_j and the other explanatory variables.

As indicated in Example 2.3 (iii), the entropy reduction $\vartheta(X_j|\boldsymbol{X}_{(-j)})$ is also the mutual information function between X_j and the other explanatory variables, $\vartheta(X_j \wedge \boldsymbol{X}_{(-j)})$. As such, $\vartheta(X_j|\boldsymbol{X}_{(-j)})$ measures functional dependency between X_j and the other explanatory variables. For the multivariate normal case the functional dependency is linear. Therefore, $\vartheta(X_j|\boldsymbol{X}_{(-j)})$ is an information measure of collinearity.

The sample version of $\vartheta(X_j|\boldsymbol{X}_{(-j)})$ is related to the traditional *variance inflation factor (VIF)* as

$$\vartheta(X_j|\boldsymbol{X}_{(-j)}) = log\ VIF_j^{1/2}. \tag{6.12}$$

The VIF is a useful and widely-used collinearity diagnostic. Traditionally, VIF_j is interpreted as the inflation factor of the variance of the sampling distribution of the least square estimate b_j, as compared with the case of orthogonal regressors. It is also interpreted as a transformation of the multiple correlation R_j^2. The relation (6.12) gives an information theoretic interpretation to VIF.

6.4 Principal Component Regression

Principal Component Regression (PCR) refers to selecting a subset of the transformed regressors in the reparametrized model (6.1) and estimating $\boldsymbol{\beta}$ based on the reduced model. The purpose of PCR is to reduce the effects of collinearity on the regression coefficients.

Let Q be a subset of the index set $\{1, \cdots, p\}$ containing q elements. Let Γ_Q denote the submatrix containing the q eigenvectors $\boldsymbol{\gamma}_j$, $j \in Q$ of $X'X$. Then the model (6.1) may be reduced as

$$\begin{aligned} \boldsymbol{y} &= (X\Gamma)(\Gamma'\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \\ &= (X\Gamma_Q)(\Gamma_Q'\boldsymbol{\beta}) + (X\Gamma_{\bar{Q}})(\Gamma_{\bar{Q}}'\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \\ &= W_Q\boldsymbol{\alpha}_Q + \boldsymbol{\varepsilon}_Q, \end{aligned} \tag{6.13}$$

where $\bar{Q}$ is the complement of Q in the set of the first p integer; W_Q and $\boldsymbol{\alpha}_Q$ contain W_j and α_j, $j \in Q$, respectively. The error term in (6.13) is defined by $\boldsymbol{\varepsilon}_Q = W_{\bar{Q}}\boldsymbol{\alpha}_{\bar{Q}} + \boldsymbol{\varepsilon}$. If the specification of the full model is correct, then for $q < p$, $\boldsymbol{\varepsilon}_Q$ does not have the covariance structure $\sigma^2\boldsymbol{I}_n$. In the sampling theory approach $\boldsymbol{\alpha}_{\bar{Q}}$ is set to zero which contradicts the full model specification. In Bayesian analysis the issue is of no concern when the prior expected value is $E(\boldsymbol{\beta}_j) = \boldsymbol{0}$; see Soofi (1988)for details.

Given an estimate $\tilde{\boldsymbol{\alpha}}_Q$, a PCR estimate of $\boldsymbol{\beta}$ is obtained by $\tilde{\boldsymbol{\beta}}_Q = \Gamma_Q\tilde{\boldsymbol{\alpha}}_Q$. The Bayes PCR estimate of $\boldsymbol{\beta}$ based on the ME normal likelihood and the ME normal prior $\pi^*(\boldsymbol{\beta}|\tau^2) = N(\boldsymbol{0}, \tau^2\boldsymbol{I}_p)$ is given by

$$\tilde{\boldsymbol{\beta}}_Q(\phi_Q) = \Gamma_Q[\Lambda_Q + \phi_Q\boldsymbol{I}_q]^{-1}\Gamma_Q'X'\boldsymbol{y}, \tag{6.14}$$

where ϕ_Q is the precision ratio for the reduced model and $\Lambda_Q = W_Q'W_Q$ which is the submatrix of (6.2) with diagonal elements λ_j, $j \in Q$.

The PCR estimate (6.14) is a general class representation of several well-known regression estimates. For $q = p$, (6.14) gives the Bayes estimate $\boldsymbol{b}(\phi, \boldsymbol{0})$ which is also the ridge regression estimate $\boldsymbol{b}(\phi)$. When $\phi_Q = 0$ and $q = p$, (6.14) gives the ordinary least square estimate $\boldsymbol{b}$ which is the posterior mean under uniform prior. The traditional PCR estimate is found by letting $\phi = 0$.

The main issue in PCR is selection of Q. The information functions (6.4) and (6.10) can be used to measure the amount of information about $\boldsymbol{\beta}$ retained in a reduced model. In the previous sections, the collinearity diagnostics were developed as if the regression error variance σ^2 and the prior variance τ^2 were known. In practice these quantities are estimated for computing the information functions.

Let $\tilde{\sigma}_Q^2$ be an estimate obtained using the reduced model (6.13). Then the amount of information (6.4) retained in the reduced models about $\boldsymbol{\beta}$ are compared according to

$$\begin{aligned} \tilde{I}_Q \equiv 2\tilde{I}_{W_Q}(\boldsymbol{\alpha}_Q|\tilde{\sigma}_Q^2) &= 2\tilde{I}_{W_Q}(\Gamma'_Q\boldsymbol{\beta}|\tilde{\sigma}_Q^2) \\ &= \sum_{j\in Q} log\left(\frac{\lambda_j}{\tilde{\sigma}_Q^2}\right) - log(2\pi e). \end{aligned} \tag{6.15}$$

Similarly, using an estimate $\tilde{\tau}^2$, the amount of information (6.10) retained in the reduced models about $\boldsymbol{\beta}$ are compared according to

$$\begin{aligned} \tilde{\vartheta}_Q \equiv 2\tilde{\vartheta}_{W_Q}(\boldsymbol{\alpha}_Q|\tilde{\phi}_Q) &= 2\tilde{\vartheta}_{W_Q}(\Gamma'_Q\boldsymbol{\beta}|\tilde{\phi}_Q) \\ &= \sum_{j\in Q} log(1+\tilde{\phi}^{-1}\lambda_j) \\ &= \sum_{j\in Q} log\left(1+\frac{\lambda_j}{\tilde{\sigma}_Q^2}\,\tilde{\tau}^2\right). \end{aligned} \tag{6.16}$$

The information criteria (6.15) and (6.16) compare models based on the relative informational value of W_Q within the set of regressors $W_1, \cdots, W_p$ as measured by the eigenvalues λ_j, $j \in Q$, and the precision of the model as estimated by $1/\tilde{\sigma}_Q^2$.

The merits of the information functions for PCR are best seen in the case of simple regression models that include a single transformed variable W_j; i.e., $Q = \{j\}$. The components of information (6.4) about the parameters $\alpha_j = \boldsymbol{\gamma}'_j\boldsymbol{\beta}$ are compared according to

$$\tilde{I}_j = 2\tilde{I}_{W_j}(\gamma'_j\boldsymbol{\beta}|\tilde{\sigma}_j^2) = log\left(\frac{\lambda_j}{\tilde{\sigma}_j^2}\right) - log(2\pi e), \quad j = 1, \cdots, p, \tag{6.17}$$

where $\tilde{\sigma}_j^2$ is the error variance for the estimated simple regression.

The components of sample information (6.10) are compared according to

$$\begin{aligned}\tilde{\vartheta}_j = 2\tilde{\vartheta}_{W_j}(\gamma_j'\beta|\tilde{\phi}_j) &= log(1+\tilde{\phi}_j^{-1}\lambda_j)\\ &= log\left(1+\frac{\lambda_j}{\tilde{\sigma}_j^2}\,\tilde{\tau}^2\right) \qquad j=1,\cdots,p. \qquad (6.18)\end{aligned}$$

Both information functions (6.17) and (6.18) compare components based on the ratio $\lambda_j/\tilde{\sigma}_j^2$. The eigenvalue signifies the relative informational value of W_j within the set of regressors $W_1,\cdots,W_p$, and the error variance indicates the relative strength of the relationship between the W_j and the dependent variable. These information functions favor components that are strong on the balance of these two features.

6.5 MDI Selection of Prior (Ridge) Parameter

In the presence of severe collinearity, the least square procedure often produces meaningless estimates for regression coefficients. The signs and/or the magnitudes of the least square estimates or of some functions of the estimated coefficients may not be meaningful. If the problem is due to collinearity, it should be corrected by reducing the effects of the collinearity in estimation. For example, when the regression matrix is orthogonal, the signs of the least square estimates correspond to the signs of the simple correlation coefficients between the regressors and the dependent variable. In the case of severe collinearity the signs of the least square coefficients may differ from the orthogonal case. Thus reduction of the extent of the collinearity should correct the problem. Since the least square estimates do not satisfy some constraints that the regression coefficients should satisfy, it should be corrected.

Consider the family of estimates constructed by the linear transforms of the least square estimate

$$\mathcal{D} = \{\boldsymbol{b}_D = \Gamma' D \Gamma \boldsymbol{b} : D = diag[d_1,\cdots,d_p],\ d_j > 0,\ j=1,\cdots,p\},$$

where Γ is the matrix of the eigenvectors of $X'X$.

The elements of the diagonal matrix $d_j,\ j=1,\cdots,p$ are the *altering coefficients*. Their role is more directly seen in estimation of the coefficient $\boldsymbol{\alpha}$ of the reparametrized model (6.1). The estimate of $\boldsymbol{\alpha}$ corresponding to $\boldsymbol{b}_D$ is given by the simpler linear transform

$$\boldsymbol{a}_D = D\boldsymbol{a},$$

where $\boldsymbol{a} = \Gamma\boldsymbol{b}$ is the least square estimate of $\boldsymbol{\alpha}$.

A well-known subfamily in $\mathcal{D}$ is defined by $\boldsymbol{b}_\Phi = \Gamma'\boldsymbol{a}_\Phi$ with

$$\boldsymbol{a}_\Phi = (\Lambda + \Phi)^{-1}\Lambda\boldsymbol{a} \tag{6.19}$$

where Λ is the diagonal matrix defined in (6.2) and Φ is a diagonal matrix with diagonal elements $\phi_j \geq 0,\ j = 1, \cdots, p$. The altering coefficients in (6.19) are

$$0 < d_j = \frac{\lambda_j}{\lambda_j + \phi_j} \leq 1, \quad j = 1, \cdots, p.$$

For $\phi_j = \phi,\ j = 1, \cdots, p$, (6.19) gives the Bayes estimate $\boldsymbol{b}(\phi, \boldsymbol{0})$ shown in (6.7) which is also referred to as the ordinary ridge estimate. The posterior mean based on the normal likelihood and the normal prior $\pi^*(\boldsymbol{\alpha}) = N(\boldsymbol{0}, \Psi),\ \Psi = diag[\tau_1^2, \cdots, \tau_p^2]$ is in the form of (6.19) with $\phi_j = \sigma^2/\tau_j^2$. In the ridge regression,(6.19) is referred to as the generalized ridge estimate (Hoerl and Kennard 1970).

When the least square is transformed in order to circumvent the ill-effects of collinearity, then it is natural to seek the minimal amount of alteration required. The alteration is considered as a perturbation of a distribution associated with the least square estimate for the inferential purposes. In Bayesian analysis, $\boldsymbol{b}$ is the Bayes estimate under the noninformative prior and the posterior distribution is altered with the use of prior. Thus the search is for identifying the minimum prior precision τ_j^2 required for an adequate estimation of $\boldsymbol{\beta}$ (Soofi and Soofi 1989). In the sampling theory approach, the sampling distribution is perturbed. Thus the problem is to select, for example, a ridge procedure that gives adequate parameter estimates with minimal perturbation (Soofi and Gokhale 1991b).

More formally, suppose that the regression coefficient is constrained as $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathcal{B}$ is a subset of $\Re^p$. Then $\boldsymbol{b}_D^* \in \mathcal{D}$ is chosen such that:

(i) $\boldsymbol{b}_D^* \in \mathcal{B}$;

(ii) $K(\boldsymbol{b}_D^* : \boldsymbol{b}) \leq K(\boldsymbol{b}_D : \boldsymbol{b})$ for all $\boldsymbol{b} \in \mathcal{D}$, where $K(\boldsymbol{b}_D : \boldsymbol{b})$ is the discrimination information function between the distributions associated with the two estimates.

For the case of normal ME likelihood and the normal ME prior, the discrimination information function is:

$$
\begin{aligned}
K(\boldsymbol{b}_D : \boldsymbol{b}|\boldsymbol{b}, \sigma^2) &= K(\boldsymbol{a}_D : \boldsymbol{a}|\boldsymbol{a}, \sigma^2) \\
&= \frac{1}{2}\left[TrD^2 + log|D^2| - p\right] + \frac{\boldsymbol{a}'(D - \boldsymbol{I}_p)\Lambda(D - \boldsymbol{I}_p)\boldsymbol{a}}{2\sigma^2}.
\end{aligned}
$$

$K(\boldsymbol{a}_D : \boldsymbol{a}|\boldsymbol{a}, \sigma^2)$ is a convex function of the altering coefficients with the global minimum at $D = \boldsymbol{I}_p$. Thus, it may minimized with respect to the altering coefficients. In practice σ^2 are estimated and the minimization is iterative. For the case of (6.19), the minimization is with respect to Φ. In some applications the formal minimization may be replaced with simpler search methods. For a Bayesian application see Soofi and Soofi (1989) and for a ridge analysis see Soofi and Gokhale (1991b). In the sampling theory approach $K(\boldsymbol{a}_D : \boldsymbol{a}|\boldsymbol{a}, \hat{\sigma}^2)$ is the MLE estimate of the discrimination information function between the distributions of the ridge and the least square estimates.

7 Influence of Observations on Information

In this section I present diagnostics for measuring influence of an observation on the distributions associated with the regression coefficients and on the predictive distribution.

Consider the case of the noninformative prior for the coefficients of the normal ME regression model (3.4). The influence diagnostics for this case are interpretable in terms of the ordinary least regression. The extension to the informative priors may be developed similarly.

Let X_{-i} and $\boldsymbol{y}_{-i}$ denote the data with the ith observation deleted. Then the change in the amount of uncertainty in predicting a value of $\boldsymbol{\beta}$ due to the presence and absence of the ith observation is given by the posterior entropy difference

$$
\begin{aligned}
\Delta H_i &= H_{X_{-i}}(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2) - H_X(\boldsymbol{\beta}|\boldsymbol{y}_{-i}, \sigma^2) \\
&= log\left(\frac{|X'X|}{|X'_{-i}X_{-i}|}\right)^{1/2} \\
&= -log(1 - h_{ii})^{1/2} \geq 0,
\end{aligned}
$$

where h_{ii} is the ith diagonal element of $X_{-i}(X'_{-i}X_{-i})^{-1}X_{-i}$; see Poston (1995) for the proof of the last equality. It is well-known that $0 \leq h_{ii} \leq 1$, thus Δ_i is well defined. The last inequality indicates that the an observation reduces uncertainty, hence is informative. The influence is negligible when $h_{ii} \approx 1$.

ΔH_i measures the influence of an observation on the extent of the collinearity. Another measure of influence on the collinearity is obtained using the change in the information number of the regression matrix

$$\Delta_i = \Delta(X_{-i}) - \Delta(X) = log\left(\frac{\kappa(X_{-i})}{\kappa(X)}\right)^{1/2}$$

where $\kappa(X)$ is the condition number of X.

The discrimination information function between the two posterior distributions is

$$\begin{aligned} 2K(\pi_{X_{-i}} : \pi_X|\boldsymbol{y}, \sigma^2) &= \left[Tr(X'X)(X_{-i}'X_{-i})^{-1} - log|X'X(X_{-i}'X_{-i})^{-1}| - p\right] \\ &\quad + \frac{(\boldsymbol{b} - \boldsymbol{b}_{-i})'X'X(\boldsymbol{b} - \boldsymbol{b}_{-i})}{\sigma^2}. \end{aligned}$$

The second term is the influence on the posterior mean which is the least square estimate. It may also be written in terms of the fitted values of $\boldsymbol{y}$. This term when the error variance is estimated by the regression mean square error equals to a multiple of the Cook distance. The discrimination information is more comprehensive then the traditional diagnostics because of the first term which measure the influence of the ith observation on the posterior variance (or variance of the sampling distribution).

Johnson and Geisser (1983) developed diagnostics for assessing the influence of observations on predictive distribution of n new observations $\boldsymbol{y}_N[X]$ corresponding to the regression matrix X. The predictive influence of a subset of observations is measured by the following discrimination information functions predictive densities: $2K[f(\boldsymbol{y}_N[X]||\boldsymbol{y}_{-s}) : f(\boldsymbol{y}_N[X]||\boldsymbol{y})]$ or $2K[f(\boldsymbol{y}_N[X]||\boldsymbol{y}) : f(\boldsymbol{y}_N[X]||\boldsymbol{y}_{-s})]$, where $\boldsymbol{y}_{-s}$ denotes the data exclusive of the subset under consideration.

For the normal ME model (3.4) with noninformative prior, the conditional predictive density is

$$f(\boldsymbol{y}_N[X]||\boldsymbol{y}, \sigma^2) = \int f^*(y|\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2)d\boldsymbol{\beta} = N(X\boldsymbol{b}, V\sigma^2),$$

where $V = X(X'X)^{-1}X'$. Similarly, the conditional predictive density when the ith observation is deleted is found to be $f(\boldsymbol{y}_N[X]||\boldsymbol{y}_{-i}, \sigma^2) = N(X\boldsymbol{b}_{-i}, V_i\sigma^2)$ where $V_i = X(X_{-i}'X_{-i})^{-1}X'$.

Then an influence measure of the ith observation is given by

$$2K[f(y_N[X]|y_{-i}) : f(y_N[X]|y)] = \; = \; \left[Tr(VV_i^{-1}) - log|VV_i^{-1}| - p\right] \\ +\frac{(b-b_{-i})'X'X(b-b_{-i})}{\sigma^2}.$$

The first term quantifies the influence on the predictive covariance. The last term, when σ^2 estimated by the mean square error, is proportional to the Cook distance.

Johnson and Geisser (1983) also developed influence diagnostics for the case of unknown variance. They used Jeffreys prior which gives the multivariate t distribution (4.19) for the predictive density. Since the discrimination information function between two t densities does not have a closed form, the influence diagnostics are developed based on an approximation. These authors extends their results to the case of normal-gamma prior (5.15). Carlin and Polson (1991) extended this line of work to case of nonlinear models.

8 References

Abel, P. S. and N.D. Singpurwalla (1994) "To Survive or to Fail: That is the Question", *The American Statistician*, 48, 18-21.

Akaike, H. (1973) "Information Theory and an Extension of the Maximum Likelihood Principle", *2nd International Symposium on Information Theory*, 267-281.

Ash, R. B. (1965) *Information Theory*, N.Y.: Dover.

Bernardo, J. M. (1979a) "Expected Information as Expected Utility", *The Annals of Statistics*, 7, 686-690.

Bernardo, J. M. (1979b) "Reference Prior Distributions for Bayesian Inference", *Journal of the Royal Statistical Society, Ser. B*, 41, 113-147.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980) *Regression Diagnostics: Identifying Influential Observations and Sources of Collinearity*, N.Y.: Wiley.

Bozdogan, H. (1990) "On the Information-Based Measure of Covariance Complexity and Its Application to the Evaluation of Multivariate Linear Models", *Communications in Statistics, Part A–Theory and Methods*, 19, 221-278.

Bozdogan, H. and M. A. Haughton (1995) "Informational Complexity Criteria for Regression Models", *Statistica Sinica*, to appear.

Brockett, P. L. (1991) "Information-Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications", *Transactions of the Society of Actuaries*, 43, 73-135.

Carlin, B. P. and N. G. Polson (1991) "An Expected Utility Approach to Influence Diagnostics", *Journal of the American Statistical Association*, 87, 1013-1021.

Carota, C., G. Parmigiani, and N. G. Polson (1996) "Diagnostic Measures for Model Criticism", *Journal of the American Statistical Association*, 91, to appear.

Cover, T. M. and J. A. Thomas (1991) *elements of Information Theory*, N.Y.: Wiley.

Csiszar, I. (1991) "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference In Linear Inverse Problems", *The Annals of Statistics*, 19, 2032-2066.

Donoho, D.L., I.M. Johnstone, J.C. Joch, and A.S. Stern (1992) "Maximum Entropy and Nearly Black Object", *Journal of the Royal Statistical Society, Ser. B*, 54, 41-81.

Ebrahimi, N. and E. S. Soofi (1996) "On the Equivalence of Entropy and Variance Orderings", under review.

Fisher, R. A. (1921) "On Mathematical Foundations of Theoretical Statistics", *Philosophical Transactions of the Royal Society of London, Ser. A*, 222, 309-368.

Geisser, S. (1993) *Predictive Inference: An Introduction*, N.Y.: Chapman and Hall.

Ghosh, M. and M. C. Yang (1988) "Simultaneous Estimation of Poisson Means Under Entropy Loss", *The Annals of Statistics*, 16, 278-291.

Goel, P.K. (1983) "Information Measures and Bayesian Hierarchical Models", *Journal of the American Statistical Association*, 78, 408-410.

Goel, P.K. and M. H. DeGroot (1979) "Comparison of Experiments and Information Measures", *The Annals of Statistics*, 7, 1066-1077.

Gokhale, P.K. and S. Kullback (1978) *The Information in Contingency Tables*, N. Y.: Marcel Dekker.

Gull, S. F. (1989) "Developments in Maximum Entropy Data Analysis", in *Maximum Entropy and Bayesian Statistics*, ed. J. Skilling, Boston: Kulwer.

Gull, S. F., and G. J. Daniell (1978) "Image Reconstruction from Incomplete and Noisy Data", *Nature*, 272, 686-690.

Ibrahim, J. and P. Laud (1994) "A Predictive Approach to the Analysis of Designed Experiment", *Journal of the American Statistical Association*, 89, 309-319.

James, W. and C. Stein (1961) "Estimation With Quadratic Loss Function", *Proceedings of the Fourth Berkeley Symposium*, 1, 361-375, Berkeley: UC Press.

Jaynes, E.T. (1982) "Prior Probabilities", *IEEE Transactions on Systems Science and Cybernetics*, SSC-4, 227-241.

Jaynes, E.T. (1968) "On the Rationale of Maximum-Entropy Methods," *Proceedings of IEEE*, 70, 939-952.

Jaynes, E.T. (1957) "Information Theory and Statistical Mechanics", *Physical Review*, 106, 620-630.

Joe, H. (1989) "Relative Entropy Measures of Multivariate Dependence", *Journal of the American Statistical Association*, 84, 157-164.

Johnson, W. and S. Geisser (1983) "A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis", *Journal of the American Statistical Association*, 78, 137-144.

Haff, L. R. (1980) "Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix", *The Annals of Statistics*, 8, 586-597.

Hill, S. D. and J. C. Spall (1987) "Noninformative Bayesian Priors for Large Samples Based on Shannon Information", *Proceedings of the IEEE Conference on Decision and Control*, 1690-1693.

Hoerl, A. E. and R. W. Kennard (1970) "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, 12, 55-67.

Kapur, J. N. (1989) *Maximum Entropy Models in Science and Engineering*, N.Y.: Wiley.

Keyes, T. K., and M. S. Levy (1996) "Goodness of Prediction Fit for Multivariate Linear Models", *Journal of the American Statistical Association*, 91, 191-197.

Kullback, S. (1959) *Information Theory and Statistics*, N.Y.: Wiley (reprinted in 1968 by Dover).

Kullback, S (1954) "Certain Inequalities in Information Theory and the Cramer-Rao Inequality", *The Annals of Mathematical Statistics*, 25, 745-751.

Kullback, S. and R. A. Leibler (1951) "On Information and Sufficiency", *The Annals of Mathematical Statistics*, 22, 79-86.

Kullback, S. and H. M. Rosenblatt (1957) "On Analysis of multiple Regression in k Categories", *Biometrika*, 44, 67-83.

Lange, K. L., R.J.A. Little, and J. M. G. Taylor (1989) "Robust Statistical Modeling Using the t Distribution", *Journal of the American Statistical Association*, 84, 881-896.

Leamer, E. E. (1979) "Information Criteria for the Choice of Regression Models, A Comment", *Econometrica*, 47, 507-510.

Lehmann, E. L. (1983) *Theory of Point Estimation*, N.Y.: Wiley.

Levy, M. S. and S. K. Perng (1986) "An Optimal Prediction Function for the Normal Linear Model", *Journal of the American Statistical Association*, 81, 196-198.

Lindley, D. V. (1961) "The Use of Prior Probability Distributions in Statistical Inference and Decision", *Proceedings of the Fourth Berkeley Symposium*, 1, 436-468, Berkeley: UC Press.

Lindley, D. V. (1957) "Binomial Sampling Schemes and the Concept of Information", *Biometrika*, 44, 179-186.

Lindley, D. V. (1956) "On a Measure of Information Provided by an Experiment", *The Annals of Mathematical Statistics*, 27, 986-1005.

Maasoumi, E. (1993) "A Compendium to Information Theory in Economics and Econometrics", *Economic Reviews*, 12(2), 137-181.

Meisner, J. F. (1980) "Maximum Entropy Regressions", *Economics Letters*, 5, 251-255.

Pesaran, M. H. (1987) "Global and Partial Non-Nested Hypotheses and Asymptotic Local Power", *Econometric Theory*, 3, 69-97.

Poskitt, D.S. (1987) "Precision, Complexity and Bayesian Model Selection", *Journal of Royal Statistical Society, Ser. B.*, 49, 99-208.

Poston, W.L. (1995) "Optimal Subset Selection Methods", *Tech. Rep. No. 117, Center for Computational Statistics*, George Mason University.

Rissanen, J. (1987a) "Stochastic Complexity", *Journal of Royal Statistical Society, Ser. B*, 49, 3, 223-265.

Rissanen, J. (1987b) "Stochastic Complexity and the MDL Principle", *Econometric Reviews*, 6, 1, 85-102.

Rissanen, J. (1986) "Stochastic Complexity and Modeling", *The Annals of Statistics*, 14, 1080-1100.

Ryu, H. K. (1993) "Maximum Entropy Estimation of Destiny and Regression Functions", *Journal of Econometrics*, 56, 397-440.

Sawa, T. (1978) "Information Criteria for Discriminating Among Alternative Regression Models", *Econometrica*, 46, 1273-1292.

Shannon, C. E. (1948) "A Mathematical Theory of Communication", *Bell System Technical Journal*, 27, 379-423.

Shore, J. E. and R. W. Johnson (1980) "Axiomatic Derivation of the Principle of Maximum Entropy and Principle of Minimum Cross-Entropy", *IEEE Transactions on Information Theory*, IT-26, 26-37.

Skilling J., and R. K. Bryan (1984) "Maximum Entropy Image Reconstruction: General Algorithm", Monthly Notes of Royal Astronomic Society, 211, 111-124.

Spall, J. C. and S. D. Hill (1990) "Least-Informative Bayesian Prior Distributions for Finite Samples Based on Information Theory", *IEEE Transactions on Automatic Control*, 35, 580-583.

Soofi, E. S. (1996) "Information Theory and Bayesian Statistics", in *Bayesian Analysis of Statistics and and Econometrics*, eds. D. Berry, K. Chaloner, and J. Geweke, N. Y.: Wiley.

Soofi, E. S. (1994) "Capturing the intangible concept of Information", *Journal of the American Statistical Association*, 89, 1243-1254.

Soofi, E. S. (1992) "A Generalizable Formulation of Conditional Logit With Diagnostics", *Journal of the American Statistical Association*, 87, 412-816.

Soofi, E. S. (1990) "Effects of Collinearity on Information About Regression Coefficients", *Journal of Econometrics*, 43, 255-274.

Soofi, E. S. (1988) "Principal Component Regression Under Exchangeability", *Communications in Statistics, Part A–Theory and Methods*, 17, 1717-1733.

Soofi, E. S. (1985) "Information Theoretic Approach to Regression", Ph.D. Dissertation, University of California, Riverside, Department of Statistics.

Soofi, E. S., N. Ebrahimi, and M. Habibullah (1995) "Information Distinguishability with Application to Analysis of Failure Data", Journal of the American Statistical Association, 90, 657-668.

Soofi, E. S. and D. V. Gokhale (1991a) "Minimum Discrimination Information Estimator of the Mean With Known Coefficient of Variation", *Computational Statistics and Data Analysis*, 11, 165-177.

Soofi, E. S. and D. V. Gokhale (1991b) "An Information Criterion for Normal Regression Estimation", *Statistics and Probability Letters*, 11, 111-117.

Soofi, E. S. and A. S. Soofi (1989) "A Bayesian Analysis of Collinear Data: The Case of Translog Function", in *American Statistical Association Proceedings of Business and Economic Statistics Section*, 321-326.

Stone, M. (1958) "Application of a Measure of Information to the design and Comparison of Regression Experiments", *The Annals of Mathematical Statistics*, 29, 55-70.

Theil, H. (1987) "How Many Bits of Information Does an Independent Variable Yield in a Multiple Regression?", *Statistics and Probability Letters*, 6, 107-108.

Theil, H. (1982) "Some Recent and New Results on the Maximum Entropy Distribution", *Statistics and Probability Letters*, 1, 17-22.

Theil, H. (1967) *Economics and Information*, Amsterdam: North-Holland

Theil, H. and C. Chung (1988) "Information-Theoretic Measures of Fit for Univariate and Multivariate Linear Regressions", *The American Statistician*, 42, 249-252.

Theil, H. and K. Laitinen (1980) "Singular Moment Matrices in Applied Econometrics", in *Multivariate Analysis* Vol. V., ed. P. R. Krishnaiah, 629-645, Amsterdam: North-Holland.

Van Campenhount, J. M. and T. M. Cover (1981) "Maximum Entropy and Conditional Probability", *IEEE Transactions on Information Theory*, IT-27, 483-489.

Vinod, H. D. (1982) "Maximum Entropy Measurement Error Estimates of Singular Covariance Matrices in Undersized Samples", *Journal of Econometrics*, 20, 163-174.

Young, A. S. (1987), "On the Information Criterion for Selecting Regressors", *Metrika*, 34, 185-194.

Zellner, A. (1994) "Bayesian Method of Moments/Instrumental Variable (BMOM/IV) Analysis of mean and regression", (Invited paper presented at ISBA2 Meeting, Alicante, Spain), H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago.

Zellner, A. (1991) "Bayesian Methods and Entropy in Economics and Econometrics", in *Maximum Entropy and Bayesian Methods*, eds. W. T. Grandy, Jr. and L. H. Schick, 17-31, Netherlands: Kulwer.

Zellner, A. (1982) "On Assessing Distributions and Bayesian Regression With g-Prior", in *Bayesian Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, 233-243, Amsterdam: North-Holland.

Zellner, A. (1976) "Bayesian and Non-Bayesian Analysis of Regression Model with Multivariate Student-t Error Terms", Journal of the American Statistical Association, 71, 400-405.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, N.Y.: Wiley (reprinted in 1987 by Krieger, Malabar, Florida).

SF 298 MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
April 1996

3. REPORT TYPE AND DATES COVERED
Technical

4. TITLE AND SUBTITLE
Information Theoretic Regression Methods

5. FUNDING NUMBERS
DAAH04-94-G-0267

6. AUTHOR(S)
Ehsan Soofi

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Center for Computational Statistics
George Mason University
Fairfax, VA 22030

8. PERFORMING ORGANIZATION REPORT NUMBER
#126

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 32850.14-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Since the publication of the seminal note, Kullback and Leiber (1951), there has been continual endeavor in statistics and related fields to explicate the existing statistical methods and to develop new methods based on the logarithmic information of SHannon (1948). There are many fine collections of information-theoretic methodologies and their applications to the related fields sych as Kullback (1954, 1959), Lindley (1956), Jaynes (1957, 1968, 1982), Theil (1967), Akaike (1973), Gokhale and Kullback (1978), Shore and Johnson (1980), Kapur (1989), Brockett (1991), Cover and Thomas (1991), Csiszar (1991), Zellner (1991), Maasoumi (1993) and Soofi (1994). During the last four decades numerous information theoretic regression methods have been developed. Kullback and Rosenblatt (1957) pioneered the information theoretic approach to regression by explicating the usual regression qyantities sych as sums of squares and F-ratios in terms of information functions. We have now information theoretic methods for model and predictive density derivation, parameter estimation and testing, model selection, collinearity analysis, and influential observation detection which can be used in sampling theory and Bayesian regression analyses. The purpose of this paper is to integrate the existing entropy-based methods in a single framework, to explire their interrelationships, to elaborate on information theoretic interpretations of the existing entropy-based diagnostics and to present information theoretic interpretations for some traditional diagnostics.

14. SUBJECT TERMS
Information theorty, model selection, entropy, diagnostics

15. NUMBER IF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as; prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NORFORN, REL, ITAR).

DOD	- See DoDD 4230.25, "Distribution Statements on Technical Documents."
DOE	- See authorities.
NASA	- See Handbook NHB 2200.2.
NTIS	- Leave blank.

Block 12b. Distribution Code.

DOD	- Leave blank
DOE	- Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports
NASA	- Leave blank.
NTIS	- Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only).*

Block 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.